人工智能作为全球风险中的积极和负面因素

埃利泽·尤德科夫斯基
(yudkowsky@singinst.org)

已于尼克·博斯特罗姆和米兰·斯科维克合编的《全球灾难风险》中 出版
2006 年 8 月 31 日稿

人工智能奇点研究所
加州 帕洛阿尔托

简介

迄今为止人工智能最大的危险是人们过早下结论认为自己理解人工智能。当然，这个问题并非仅限于人工智能领域。 雅克·莫诺 写道：“进化论令人感到好奇的一点是人人都认为自己理解它。”（莫诺 1974）我的父亲是一位物理学家，他抱怨人们自己编造物理学的理论，并想知道人们为什么不自己编造化学理论。（答案是：他们也那么做。）尽管如此，这个问题在人工智能方面显得尤为突出。人工智能领域一向以目标远大而无法兑现着称。大多数观察者总结认为人工智能是个难题；事实上它也是。但是人工智能的**窘境**并非源自它的困难性。想从氢原子造出一个星球来也很难，但星体天文学领域却未曾因为许诺造出星星却没兑现而有坏名声。由此可见关键问题并**不**是人工智能本身有多困难，而是人们太容易地高估自己对人工智能问题的了解程度。

正如我在《全球灾难风险》的另一章中指出的，“认知偏差潜在地影响着我们对全球风险的判断”，开场白中我提到很少有人会故意选择消灭世界；因此非常令人担忧的情形是地球由于**失误**而被毁灭。极少有人会在清楚明白按下某个按钮会引起全球灾难的前提下仍那么做。但是如果人们非常容易确信某个按钮的功能而实际上它所造成的后果并非如他们所想，那样就确实值得警觉。

相比认知偏差而言，要写关于人工智能的全球风险要难得多。认知偏差是成熟的科学，只要提及相关文献就可以。人工智能是新兴的科学；它属于前沿，而不是教科书。并且由于在后面章节将会提到的原因，关于人工智能的**全球灾难风险**的话题，几乎找不到任何技术文献上的相关讨论。我必须从我自己的观点中分析问题；给出我自己的结论，并在有限的空间内尽己所能进行论证。并非是我忽视了对该话题的主流工作的引用，而是，就我最大认识所能看来，没有主流的工作成果可以被应用 （截至 2006 年 1 月）。

忽略人工智能的做法很诱人，因为在本书提到的所有全球风险中，人工智能是最难讨论的。我们不能像对小行星撞击地球那样，通过查阅实际的统计学资料赋予一个小的年度发生概率值。我们不能像对拟议的物理灾害那样运用微积分从一个

非常精确的、经过证实的模型中排除事件发生的可能性或是对它们的发生概率置一个无穷小的上限。但这只能使人工智能所引起的灾难更多令人担忧，而非更少。

研究发现许多认知偏差的效应随着时间的压力，认知忙碌程度，或信息稀缺度而**增加**。也就是说，**越难分析的偏差**，避免和减少它就越为重要。因此我**强烈推荐**在继续下文之前首先阅读“认知偏差潜在地影响着我们对全球风险的判断”一章 （第 XXX-YYY 页）。

1：拟人化偏见

当一样事物在我们日常生活中足够普遍时，我们往往想当然地忘了它的存在。

设想一个具有十个必需部分的复杂生物适应系统。如果十个基因中的每一个都在基因库中以 50%的概率独立出现 – 每个基因在该物种的一半生物体中出现 – 那么，平均说来，每 1024 个生物体中只有一个会拥有完整的有功能的适应性。除非生物体一直处于低温环境中，否则羽毛并不具有进化上的优势。类似的，如果基因 B 依赖于基因 A，那么除非基因 A 能形成该**遗传**环境中的可靠表型，否则基因 B 就没有进化优势。**复杂的，相互依赖的**身体构成在一个有性繁殖物种中必然是**普遍的**；否则它不能进化。（托比与考斯曼德 1992）一只知更鸟的羽毛可能比另一只的光滑，但它们都有翅膀。自然选择，源于变异，却将它耗尽。（舒博 1984）

在每一种已知的文明中，人类体验喜悦，悲伤，厌恶，气愤，恐惧和惊讶 (布朗 1991)，而且用同样的面部表情来显示这些情绪 （艾克门与凯尔特纳 1997）。我们的头巾下都运行着同样的引擎，尽管我们肤色各异；这就是被进化心理学家称为**人类精神统一性**的原理 （托比与考斯曼德 1992）。这一观点既能被进化生物学的机制所解释，又是其所必要的。

一位人类学家不会这样兴奋地报告他新发现的一个部落：“他们吃食物! 他们呼吸空气！ 他们使用工具! 他们互相讲故事！”生活在一个只强调差异性的世界里，我们人类忘了我们之间的共性。

人类已经进化到可以模仿他人的程度 – 和我们自己同种个体相互竞争与合作。在原始社会里一个可靠的特性是你遇到的任何一个具有强大智慧的个体一定是人类。我们进化到能够通过设想体验他人的处境而对我们的同类**感同身受**；因为被模拟者和模拟者本身是相似的。毫不奇怪，人类经常会“拟人化” – 赋予非人类以人化特性。在《黑客帝国》（沃卓斯基兄弟 1999）中，假设的“人工智能”

特工史密斯一开始表现出完全的冷静和镇定，他的面部显得被动和毫无表情。但在后来审讯人类中的莫菲斯时，特工史密斯发泄了他对人类的厌恶 – 他的脸上显示出人类普遍的面部表情，厌恶。

如果你需要预测他人的想法时，询问自己的大脑活动是有益的，这是一种适应性本能。当你面对其它类型的优化程序 – 譬如， 假设你是十八世纪的神学家威廉·佩利，正观察着生命的复杂秩序并惊奇于它从何而来 – 这时拟人化对粗心的科学家来说是一个诱捕器，这个陷阱是如此粘着以至于需要一个达尔文方能从中逃脱。

有关拟人论的实验显示被试者会无意识地运用拟人化方法，而且常常与他们深思熟虑的信念相悖。在巴雷特和凯尔（1996）的一项研究中，被试者强调他们坚信上帝非拟人化的特性：即上帝可同时在多个地点出现，或能同时关注多个事件。巴雷特和凯尔给同一批被试者呈现一些故事，例如，上帝拯救溺水者的事。当被试者回答关于该故事的一些问题，或用他们自己的话来复述故事时，他们的陈述暗示上帝同一时间只在一个地方并且是顺序地而非并行地执行任务。一个支持我们论点的意外收获是，巴雷特和凯尔还对另一组被试讲了同样的故事，唯一的区别是在那里上帝被代之以名为“Uncomp”的超智能计算机。例如，为了模拟无处不在的特性，被试者被告知 Uncomp 的传感器和效应器“覆盖地球上的每平方厘米因此没有信息会被遗漏”。在该条件下被试也表现出很强的拟人化倾向，尽管比上帝那组要少很多。从我们的角度看来，这个实验的关键结果是表明在人们有意识地相信人工智能与人类不同的情形下，他们仍会可视化一些人工智能的拟人化场景 （但不如对上帝的拟人化程度那么高）。

拟人化偏差可以被归为隐伏的认知偏差一类：它在没有明确的意图下发生，无需自觉意识，即便明确了解相关的知识也无法避免。

在科幻小说的低级趣味时代，杂志封面上偶尔会描绘着一个有情感的外星怪兽 – 俗称暴眼怪物或 BEM – 夺走一个裙子被撕破的迷人的女性人类。画家似乎认为一个具有完全不同进化史的非人类的异形，会对人类的女性有性渴望。人们不会为犯那样的错误而辩解说：“所有的心智似乎都具有相同的联接方式，由此推测一个 BEM 应该会**觉得**人类的女性富有性吸引力。”可能那个画家没有去**问**一个巨虫是否**觉得**女人很有吸引力。相反，一个裙子被撕破的女人本身是**很性感的** – 那是个内在特性。犯该错误的人没有从一个多脚昆虫机器人的角度思考；他们只是注意了女人被撕破的裙子。如果裙子没被撕破，那女人也就不那么性感；BEM 也就不会抓她。（这例子说的是一个根深蒂固的，令人困惑而又极其普遍的认知偏差错误，它被 E.T.杰恩斯称为**心智投射谬误**。（杰恩斯与巴雷特绍斯特， 2003）。杰恩斯，一位贝叶斯概率理论家，创造了“心智投射谬误”一词用来指代由主观性介入而引起知识状态混乱的认知错误。例如，**神秘现象**一词暗示神秘性是该现象本身的特性。如果我对该现象无知，这其实是我自己的心智状态，而不是该现象的实在特性。

人们无需意识到他们正在进行拟人化（或甚至无需意识到他们正参与一种预测他人思想的可疑行为）来让拟人观凌驾于认知之上。当我们试图对他人的想法进行推理时，推理过程的每一步都可能被人们经验中的世故假设所污染，就像我们对空气和重力一样不加注意。你向杂志插图画家抗议道："那么巨型大公虫是否更像是会对巨型母虫产生性欲望呢？" 插图画家想了片刻后对你说："哦，即使一个昆虫异形一开始喜欢坚硬的外骨骼，当它遇到人类女性后它很快会明白她们拥有更为美丽的，柔软的肌肤。如果那些异形具有足够先进的技术，它们可以通过遗传工程自行改造以使自己喜欢柔软的皮肤而不是坚硬的外骨骼。"

这是很自然的谬误。当异形的拟人化想法被指出时，杂志插图画家试图退一步把异形的结论辩解为是它的推理过程的中性产物。也许先进的异形**会**重组它们自身（通过遗传上或其它方式）令其喜欢柔软皮肤，但他们会**想要**那么做吗？一个喜欢坚硬外壳的昆虫异形是不会希望把它自己变得喜欢柔软皮肤的 - 除非自然选择使它产生了超性别的明显人类感知。当使用长的，复杂的推理链来为某个拟人化的结论辩护时，推理过程的每一步都成为错误可能潜入的机会。

不仅如此，从结论出发寻找一条通往该结论的貌似中性的推理线索本身也是个严重的错误；这是合理化。如果是大脑自询产生了昆虫机器人追逐女人的第一瞬的心智图像，那么拟人化是那一信念的潜在原因，再多推理也无法将其改变。

对任何想要减少自己拟人化偏差的人来说，一条好建议是通过学习进化生物学，尤其是用数学方法描述的进化生物学，来练习去除该偏差。早期的生物学家常常把自然选择拟人化 - 他们认为进化会做类似他们自己做的事情；他们试图"穿上进化的外衣"来用自己的想法预测进化的效应，结果得到一派荒谬的言论。这些谬论在 60 年代晚期首先从生物学开始被**系统地**铲除，例如威廉姆斯所为（1966）。进化生物学提供了数学和案例分析以帮助消除拟人化偏差。

1.1： 智力设计空间的广度

进化强有力地保留某些结构。当依赖于已知基因的另外基因进化时，那个早期的基因是固定不变的；它一旦突变就会破坏多重关联适应性。同源异形基因 - 那些控制胚胎期身体架构发育的基因 - 告知许多其它基因该何时活化。一个同源异形基因突变会引起果蝇胚胎发育成其它部分均正常却无头的个体。正因为这样，同源异形基因被如此强有力的保留下来，以至于他们在人和果蝇中是一样的 - 他们从最后一个人和昆虫的共同祖先开始就没有被改变过。ATP 合成酶的分子机器在动物的线粒体，植物叶绿体和细菌中本质上是一样的；ATP 合成酶从 20 亿年前真核生物诞生以来就没有发生过重大改变。

而两个人工智能设计之间的差异却可能大过你和一株矮牵牛花。

"人工智能"这个术语指向一个比术语"人类"更广阔许多的**可能性空间**。当我们谈及"人工智能"时我们事实上是在谈论**一般意义上的智力**或说是一般性的优化过程。想象一下一张智力设计空间的地图。在一个角落，一个小小的圈包含了所

有的人类；稍大一点的小圈包括所有的生物体；而这张巨大地图的其余部分是**一般智力的空间**。这整张地图浮在一个更大的空间里，即所有优化过程的空间。自然选择并非刻意地创造出复杂的功能机械；进化位于**优化过程的空间**内，但在包括大脑的空间之外。

正是这个**庞大的**可能性空间使得拟人化的推理方式不再合理。

2: 预测与设计

我们不能通过查询我们自己的大脑来寻找非人类的优化过程 – 不管是暴眼怪物，自然选择，或是人工智能。那么我们从何入手呢？我们怎样预测人工智能能做什么？我故意以这样一种方式提出问题从而使它看上去十分棘手。通过这个令人望而却步的问题，我们无法预测一个**任意的**计算系统是如何实现任何输入-输出转换功能的，包括，譬如说，简单的乘法。（莱斯 1953）那么人类工程师又是如何造出可靠地实现乘法运算的计算机芯片的？那是因为人类工程师刻意使用了他们**能够**理解的设计。

拟人论使人们相信他们能够预测，在除了知道它有“智能”之外一无所知的情况下 – 拟人化可以继续生成预测，而不管你们的大脑不自觉中已把它自己穿上“智慧”的外衣。这可能是导致人工智能尴尬历史的因素之一，那些尴尬不是由人工智能的困难产生，而是由对于一个给定的人工智能设计的功能所形成错误信念的不可思议的容易性产生。

若要声明一座桥可以载重达 30 吨的车辆，土木工程师有两个武器：选择初始条件，以及设定安全界限。他们不用去推测**任一**结构是否能承载 30 吨的车辆，他们只要造出一座桥使该声明成立。尽管一个工程师能够精确计算一座桥的载重量是很了不起的，要是只能计算出一座桥**至少**可以载重 30 吨也是可以接受的- 尽管要**严格**实现这个貌似含糊的声明所需要的理论理解力，和一个精确计算所需的也差不多。

土木工程师以高标准衡量他们自己对桥的载重性的估计能力。古代炼金士衡量他们自己将铅变成黄金的能力的标准则要低得多。多少铅可以变成多少金子？什么是它转变的精确因果机制？炼金研究员**想要**把铅变成黄金的原因是显而易见的，但为什么这一系列的试剂会使铅变成金，而不是把黄金变成铅，或是把铅变成水呢？

一些早期的人工智能学者相信一个有多层阈值单元组成的人工神经网络，通过反向传播训练，可以拥有“智慧”。这种痴心妄想更类似于炼金士的想法，而不是土木工程师的。在唐纳德· 布朗的清单上魔法是人类的普遍想法（布朗 1991）；而科学不是。我们并不能**凭直觉**意识到炼金术行不通。我们不能**凭直觉**严格区分确实的知识和构思精巧的故事。我们无法**本能地**注意到一个有正面结果的预期其实是空中楼阁。

人类经过自然选择而生存下来，自然选择通过对偶然突变的必然保留而起作用。一条通往全球灾难的途径 – 如同某人按下一个对其功能有错误理解的按钮 – 就是人工智能通过一种类似自然选择的操作算法不断累加而形成，而研究者对于整个复合系统如何工作却没有深刻的认识。
尽管他们相信所产生的人工智能是友好的，他们却对产生友好行为的精确过程印象模糊，或是对友好的具体含义理解不清。正如早期人工智能学者对他们程序的智力预期是模糊而严重错误的，我们试想那些人工智能学者成功地构造了一个智能程序，却对他们程序的友好性持有模糊而严重错误的预期。

不知如何构建一个友好人工智能本身，在任何特定情况下都并非是致命的，如果你明确知道自己不知如何构建。反而是那种认为人工智能一定会对人类友好的**错误**信念蕴含着通向全球灾难的危险。

3：智能力量低估

我们倾向于关注个体差异而不是人类普遍性。因此当有人提到“智力”一词时，我们想到的是爱因斯坦，而不是全人类。

人类智力的个体差异有一个标准称谓，即斯皮尔曼的 g 亦称为**g-因子**，它是对可靠实验结果的具有争议的解释，实验结果表明不同智力测试之间相互高度关联，而且智力测试结果与现实世界中的结果如终身收入高度关联。（约翰逊 1999）斯皮尔曼的 g 是从人类之间智力的个体差异得到的统计抽提，人类作为一个**种群**远比蜥蜴的智力要高。斯皮尔曼的 g 是抽象表示出巨人族里豪米级别的高度差异。

我们不应混淆斯皮尔曼的 g 和**人类一般智慧**，即我们解决许多对其它物种而言无法理解的各式认知任务的能力。一般智力指的是物种间的差异，一种复杂的适应性，一种在所有已知文明中都能找到的人类普遍性。也许关于智力没有学术上的统一观点，但对于这件亟待解释的事物的力量和存在性却是毫无疑义的。人类拥有**某些特性**使我们登上月球。

然而“智力”一词常常会使人联想起智商 160 而食不果腹的教授和智商 120 的亿万富翁、公司首席执行官。事实上除了“聪明”外促成某人在人类社会中获得相对成功的个体能力差异还包括：热情，社交技巧，教育，音乐天赋，理性。请注意我列举的每个因素都是**认知**方面的。社交技巧存在于大脑中，而不是肝脏。撇开笑话不谈，你不会发现有许多黑猩猩成为 CEO，或学术界教授。你也不会找到哪只老鼠成为受欢迎的理性主义者，或艺术家，或诗人，或领导者，或工程师，或老练的社交家，或武术家，或作曲家。智力是人类力量的基石，它是我们获得其它技艺的源泉。

混淆普通智力与 g-因子的危险是会导致对人工智能潜在影响的大大低估。（这既

指潜在良性影响，也指潜在负面影响。）即使“超人类人工智能”或“人工超级智慧”这些词也仍然只是唤起人们脑海中固有的书呆子形象：它擅长一些传统认为与“智力”相联的认知任务比如象棋或抽象数学之类，而不是具有超越人类的说服力；或对人类社会的情境能做出远比人类更好的预测和操控；或是在制定长期战略方面超人地聪慧。如果谈到智力时我们不提爱因斯坦，那么我们是否应该想到 19 世纪政治和外交天才奥特· 冯· 俾斯麦？但那只是同样错误的一个翻版。从一个乡村笨蛋到爱因斯坦，或从一个乡村笨蛋到俾斯麦所覆盖的智力范围，只是从阿米巴虫到人类智力全范围中的一小点。

如果“智力”一词使人想起爱因斯坦而不是人类，那么说智力比不上一把手枪或说智力比不上金钱听上去也会是合理的，但这就像说手枪长在树上或老鼠会用钱一样荒谬。人类从一开始就不具备其它物种赖以生存的爪，牙，盔甲或其它优势。如果你从生物圈中其它生物的角度来观察人类，你看不出这种湿软的物种是怎么最终会给他们自己穿上装甲坦克。我们发明了战胜狮子和狼的武器。我们没有像它们一样的爪子和尖牙；我们有自己心知肚明的重要优势，这就是创造力的力量。

文基 （1993）恰当地观察到未来具有比人聪明的思维能力的是另类的物种。人工智能不是最新一期科技杂志上广告中那令人惊奇的闪亮的昂贵小器具。人工智能不属于展示医药，制造和能源进展图表中的一项。你不能随意将人工智能与未来剧的滥俗情景中的摩天大楼和飞行汽车，以及可以让你停止呼吸八小时的纳米血红细胞混为一谈。再高的摩天大楼也不可能开始做自己的工程。人类在地球上取得显赫地位也不是靠他比其它物种屏息时间更长。

低估智慧的力量所引发的灾难性后果类似于一些人造了一个按钮，而对该按钮的功能不够重视，因为他们不认为那个按钮强大到足以伤害到他们。或者，由于低估智力的力量意味着相应地低估人工智能的潜在影响，目前极少数代表人类处理生存危机的研究人员和资助者以及个人慈善家可能对人工智能将不够重视。或是更广范围内的人工智能领域将会对强大的人工智能不够重视，因此当构造一个强大人工智能成为可能的时候，实现友好性的便利工具和坚实基础还没有准备好。

不得不提的是 – 由于它也影响到生存危机 – 人工智能可能是其它生存危机的有力解决方案，而我们错误地忽略了我们存活的最大希望。了解低估人工智能潜在影响的关键是意识到其潜在的好坏影响是对等的。那就是为什么本章的标题是“人工智能作为全球风险中的正面和负面因素”，而不是“人工智能的全球风险。”人工智能的未来与全球风险的关系比那更为复杂；如果人工智能是一个纯粹的责任，问题会变得简单。

4：能力和动机

在探讨人工智能，特别是具有超人类能力的人工智能时有一个常犯的错误。有人说：“当技术足够先进时我们可以有能力制造远超人类智能的大脑。既然，很明显你能做多大的奶酪蛋糕取决于你的智力，那么一个超级智慧体可以做巨大的蛋

糕 – 有一个城市那么大的蛋糕 – 天哪，未来将充满了奶酪蛋糕！”问题是那个超级智慧体是否**想要**做大蛋糕？这个例子里，人们的关注点从**能力**一下跳到了**现实**，没有考虑必需的中间环节：**动机**。

以下的推理链，无需支持论据可独自成立，都显示出大蛋糕谬论：

- 一个足够强大的人工智能可以制服任何人类的反抗并且将人类消灭。【而且该人工智能将决定那么做。】因此我们不应该构造人工智能。
- 一个足够强大的人工智能可以开发出挽救百万人生命的新药技术。【而且该人工智能将决定那么做。】因此我们应该构造人工智能。
- 一旦计算机变得足够便宜，大多数的工作由人工智能来完成比人类更方便。一个足够强大的人工智能甚至会比我们更擅长数学，工程，音乐，艺术，和我们认为有意义的所有其它工作。【而且该人工智能将决定完成那些工作。】因此当人工智能发明后，人类将无事可做，我们要么挨饿，要么看电视。

4.1： 优化过程

以上对大蛋糕谬误的解析引发了一个内在的拟人化思想 – 即动机是可分离的想法；隐含的假设是通过把“能力”与“动机”作为相互独立的实体来谈论，我们在现实的接合处将其分割，这是一种实用的切割但同时也是一个拟人化的做法。

为了用更一般的术语来考察问题，我引入**优化过程**的概念：一个要在大的搜索空间中命中小目标从而产生连贯的现实世界效应的系统。

一个优化过程将未来引导入可能的特定区域。我在一个远方城市访问，当地的一个朋友志愿送我去机场。我对附近不熟。当我的朋友开到十字路口时，我不知道他将转向何处，不管那是一个单独的还是一系列接连的转弯。但我却能预测我朋友这些无法预测行为的**最终结果**：我们会到达机场。即使我的朋友家在城市的另一个地方，从而他会走一连串完全不同的转弯，我仍能确定地预测我们的目标。从科学上说，这难道不是一种奇怪的情况吗？我可以预测一个过程的**结果**，而无需预测该过程**中间步骤**的任何一步。我把优化过程所导向的未来区域称为该优化器的**目标**。

设想一部轿车，假如是丰田花冠，在组成这部花冠的原子的所有可能的配置中，只有无穷小的一部分可以成为一部有用的轿车。如果你随机地去组装分子，那么你要得到一部轿车得花费**许许多多**宇宙纪元的时间。设计空间的一小部分确实描绘了那些我们觉得比花冠更快，更高效，更安全的车辆。因此花冠在设计者的目标中不是**最优**的。然而无论如何，花冠确实是，**经过优化的**，因为设计者为了创造一辆可工作的轿车必须从设计空间中命中相对无限小的目标，更何况是花冠这样高质量的轿车。你不能通过随机锯断木板和按照抛硬币的结果敲钉子来建造一辆可以有效使用的四轮马车。要想命中配置空间中如此小的目标需要一个强有力

的优化过程。

“优化过程”这个概念如**意料中的一样有用**，因为理解一个优化过程的**目标**比理解它一步步的**动态过程**要容易。上面关于花冠的讨论**隐含的**假设是花冠的设计者试图要产生出一个“车辆”，一种交通工具。这个假设应该被显式说明的，它没有错，而且对于理解花冠非常有用。

4.2： 针对目标

如果忘了一般智力空间比人类这个小点要广阔许多的事实，就会被“人工智能们”将会“想要”什么这个问题诱惑。人们必须抵制诱惑而把范围扩大到所有的可能智力空间。写故事的人编织我们称之为“未来”的这片遥远而陌生的土地的故事，说未来**将会怎样**。他们进行**预言**，他们说，“人工智能的机器人军队会进攻人类”或“人工智能会发明一种治疗癌症的方法”。他们不提供从初始条件到结果之间的复杂关系的描述 – 那会使他们失去听众。但我们必须**掌握**通往未来的具体过程的相关知识，将它引导入一个令人愉快的区域，如果我们不加引导，就会使它驶入危险。

关键的挑战不是**预言**“人工智能”会动用机器人军队进攻人类，或是发明一个癌症疗法。我们的任务甚至也不是对**任意的**一个人工智能设计加以推测。我们的任务是选择一些**特定的**强大的优化过程加以实现，并且我们能够合理断定那些过程是有益的。

我**强烈呼吁**我的读者不要有这样的先入之见：认为一个完全通用的优化过程必然是友好的。自然选择不是友好的，但它既不恨你，也不会置你不顾。进化不会拟人，它不会像你一样运作。许多 60 年代早期的生物学家期望进化会完成所有美妙的事情，他们还对进化为什么要那么做寻找各种详细的理由。他们以失望告终，因为进化本身并不是从要使人类愉快的目标出发，进而通过选择压力使产生美好结果的精细方法合理化。因此自然界实际发生的事件的原由与 60 年代早期生物学家们的想法是不同的，故预测与现实产生了分歧。

一厢情愿的空想会填充细节，限制预测，因而牺牲了实际可能性。一个希望桥不会倒塌的工程师会怎么说？工程师应该声称一般情况下桥都不会塌吗？然而自然本身不会给出桥不塌的理由，是土木工程师在特定的知识指导下通过特别的选择克服了似乎不可能克服的困难。一个土木工程师从设想要造一座桥出发；在严格的理论指导下选择一种可以承载车辆的桥的设计；进而在真实世界里造出一个反应该经过仔细计算的设计的桥梁；因此那个真实世界中的桥梁结构能够承受车重。由此达到了预期结果与实际结果的和谐统一。

5：友好人工智能

如果人们知道该如何选择一个达到某种特定目标的优化过程并加以实现就太好了。

或通俗地说，如果我们知道怎样造一个美好的人工智能就太好了。

要描述迎接该挑战所需的**知识领域**，我提议使用“友好人工智能”一词。除了指代技术本身以外，“友好人工智能”还可以指代技术的**产物** – 一个源自特定动机的人工智能。在这两种情况下我提及**友好**一词时，我都将首字母大写以与直观上理解的“友好的”相区别。

我遇到的一个常见反应是人们会立即宣称友好人工智能是不可能的，因为任何足够强大的人工智能都会修改它自身以打破我们对于它友好性的限制。

首先你需要注意的错误是大蛋糕谬误。一个可以自由访问其自身源代码的人工智能，原则上说，具有修改它自己的优化目标的**能力**。但这并不意味着它有修改它自己动机的**意图**。我不会故意去吞食一个让我享受杀人乐趣的药丸，因为**当前**我不希望我们人类会灭亡。

但是，如果我试着修改自己，而犯了错误怎么办？当计算机工程师**检验**芯片有效性时 – 要检验一个有 1.55 亿个晶体管组成而事后不能打补丁的芯片可不是件容易的事 – 工程师使用人工指导的，机器执行的程序化校验。一个**公式化**的数学校验的美妙之处是验证一千万步和验证十步是同样可靠的。但是人类可不适宜亲自去审核传说中的一千万步；我们太容易犯错了。并且目前的定理证明技术还不足以聪明到自行设计并验证一整片计算机芯片 – 当前的算法在搜索空间中以指数爆炸级方式运算。人类数学家能证明的定理远比当代定理证明机器可处理的要复杂得多，而且不会引起指数爆炸。但是人类的数学是非正式和不可靠的；间或会有人发现以往已被接受的非正式证明中的错误。解决办法就是人类工程师在每个**中间**步骤上指引一个定理证明机器。人类选择下一个辅助定理，复杂的定理证明机器产生正式的证明，再由一个简单的校验器加以验证。那就是当代工程师能够制造出有 1.55 亿个独立组件的可靠机械的原因。

验证一个计算机芯片正确有效需要人的智能与计算机算法的协同合作，由于**目前**它们任何一者本身都无法独自胜任。也许一个真正的人工智能可以用类似的**兼具两种能力**的方法来修改自己的源码 – 它既能**发明**和设计不受指数爆炸困扰的大型复杂算法，而**又**能极其精确可靠地**验证**它的每一个步骤。这是使真正的人工智能在追逐目标过程中仍保持稳定状态的一条途径，即使它需要进行大量自我修改。

本文将不对上述的想法做详细展开。（相关概念可以参考 施米德胡贝 2003 年 文献）但是我们在宣称一件具有挑战性的事物不可能实现**之前**，必须详尽考察当前已有的最先进的技术 – 尤其是当结果可能会带来巨大风险时。如果在没有经过仔细研究和给创造性以一试身手的机会前，宣布某一挑战是无法战胜的，就是对人类天赋的贬低。说你**不可能**做到某事的话是武断的 – 说你不可能造出比空气重的飞行器，说你**不可能**从核反应中获得有用的能量，说你**不可能**飞向月球。说那样的话太过宽泛，把任何一个人所能想到的每一个解决问题的可能性给完全抹杀了。只要一个反例就能推翻一个广谱定论。说友好人工智能**理论上不可能**这句

话，是冒失地否定了**每一个**可能的智力设计以及**每一种**可能的优化过程 – 包括了也具有智力的人类，他们中的一些是友善的并且希望自己变得更为友善。此时此刻，有无数的模糊不清的各种理由认为友好人工智能非**人力**所及，或是即便问题是可以解决的，也没有人能在可预见到的时间内解决它，但是我们不应过早就把这个难题一笔勾销，特别是考虑到它可能的风险。

6: 技术失败与哲学失败

博斯特罗姆 （2001 年）将一个生存灾难定义为一次可以永久灭绝地球上所有智慧生命或是**所有产生智慧生命潜在可能性**的大灾难。我们可以粗略地将导致期望中的友好人工智能失败的潜在因素分为两类，**技术失败**与**哲学失败**。技术失败指的是当你造了一个人工智能而它却不按你预期的方式运作 – 你没能理解你自己代码的真正功能。哲学失败指的是试图制造错误的东西，以至于即使你成功了，仍然不能帮助他人或对人类有益。毋庸赘述，这两种错误并非互相排斥。

两者之间的边界是细微的，因为大多数哲学失败用技术知识来解释都会容易得多。理论上你应该先说明白你**想要**什么，再弄清楚**怎么**得到它，而实际上要想弄清你想要什么往往需要对可实现它的技术有深刻的理解。

6.1：一个哲学失败的例子

19 世纪晚期，许多聪慧而诚实的人提倡共产主义，他们都是出自善意。首先发明、传播、和吸收共产主义文化基因的人，从冷静的历史观看来，是理想主义者。**第一批**的共产主义者没有苏联的史例来警告他们。**在当时的情况下，没有事后之见，共产主义听上去一定是一个很好的想法。**在革命之后，当共产主义者当权并被权力腐蚀时，其它的企图也乘虚而入；但这些本身不是最早的理想主义者所期望的，不管它们的发生是多么容易被预期到。**重要的**是我们要明白大灾难的始作俑者并不一定是邪恶的，甚至也不是愚蠢**至极**。如果我们将每一个悲剧都归咎于邪恶或愚昧，我们就会审视自己，自以为既不邪恶也不愚蠢，并说道：“那永远不会发生在**我们**身上。”

基于对他们可能的革命成果经验上的判断，早期的共产主义革命家期望发生的事情是人们的生活会得到改善：劳动者不必再承受长时间的、繁重的却报酬极少的劳动。客气一点说，结果却并非是这样。但是早先的共产主义者**预期**将会发生的结果，和其它政治体系对**它们**各自偏好的政治系统的基于经验的结果预期也没有太多的不同。他们都以为结果会让人民幸福。他们错了。

现在假设有人试图编写一个“友好的”人工智能来实现共产主义，或自由主义，或无政府主义，或**其它什么政治体系**，并怀着给人类带来乌托邦的美好愿望。人们最喜爱的政治系统往往能激起阳光般灼热的积极情绪，所以对于它的提倡者来说那个提议听上去实在是个好主意。

我们可以从道德和伦理的角度来观察编程者的失败原因 – 说那是由他们的过度自信，没有考虑到自己可能犯错，拒绝考虑共产主义自始至终都可能是一个错误的可能性。但从贝叶斯决策理论的角度，对这个问题有一个技术上的补充观点。从抉择理论上看，选择共产主义是一种经验信仰与价值评判相结合的产物。其**经验**信仰就是一旦共产主义得以实现，就会有一个或一类结果：人们更加快乐，工作时间更短，获得更大的物质财富。这完全是一个基于**经验**的预测；即使快乐在某种程度上是大脑状态的一个真实属性，尽管它有点难于衡量。如果你实现了共产主义，这种结果要么发生，要么不发生。价值评判是指这个结果与当前情形相当或是更为令人满意。对于一个共产主义系统在**真实世界里实际后果**的**经验**信仰如果改变了，那么决策也会相应发生变化。

我们希望一个真正的人工智能，一个通用人工智能，有能力改变它的经验信仰。（或者它的概率世界模型等等）。如果由于某种原因查尔斯·巴贝齐在尼古拉斯·哥白尼之前诞生，计算机在望远镜之前被发明，并且当时成功地创造出了通用人工智能，那个人工智能并不会一直相信太阳绕着地球转。那个人工智能也许可以超越它的缔造者的知识误区，鉴于编写该人工智能的程序员对推理远比天文学擅长。要造一个**发现**行星运动轨迹的人工智能，程序员不用知道牛顿力学所涉及的数学知识，他只需知道贝叶斯概率定理所涉及的数学。

编写一个实现共产主义或任何其它政治系统的人工智能的可笑之处在于，你是在编写**手段**而不是**目的**。你是在编写一个一成不变的决定，在获得了关于共产主义后果的进一步的经验知识后也不对该决定重新评估。你是在给人工智能一个固定的决定而且不告诉它如何在一个更高的智慧水平上，重新计算产生那个决定的不可靠的过程。

如果我和一个比我强大的对手下棋，我将无法**准确**预测对手的行为 - 否则我就至少是和他一样强了。但我能预测最后的结果，那就是他会赢。我知道当我的对手以获胜为目标的情况下可能的一些结果，那使我推测他会获胜，尽管我不知道中间步骤。当我充分发挥创造性的时候，也即当我的行为最难预测时，对我行为的**结果**却是**最容易**预测的。（前提是你知道并理解我的目标！）如果我想要一个强过人类的棋手，我必须编写一个**搜索器**搜寻可以获胜的棋子移动方案。我不能编写一个特定的移动方案，否则编出来的棋手就不会比我高明。当我发起搜索时，我就牺牲了事先预测**正确**答案的能力。要找到一个真正完美的答案你必须牺牲你预测答案的能力，但别丢了你辨别什么是问题的能力。

直接去编写共产主义的胡涂之举，应该不会发生在一个理解决策理论的通用人工智能的程序员身上。我称它为哲学失败，但它是一个由技术知识缺乏引起的哲学失败。

6.2：一个技术失败的例子

“对于智能机器的行为我们不应制定规则加以限制，而是应该赋予智能机器情感，

让情感引导它们对行为的学习。他们应该想让我们愉快和繁荣，那是我们称之为喜爱的情感。我们可以设计智能机器以使它们最基本的，与生俱来的情感就是对人类无条件的喜爱。首先我们可以造一些相对简单的机器，它们能够学习从人类的面部表情、声音语气以及肢体语言识别出人们是快乐还是不快乐。接着我们可以将这些学习结果作为先天的感情观固化到更为复杂的智能机器中，它们会增强使我们快乐的行为而减弱使我们不快乐的行为。机器可以学习算法对未来作近似地预测，例如投资者当前使用学习机预测证券价格。因此我们可以编程使智能机器学习预测未来人类的快乐程度，并用这些预测来作为它们的情感价值。”

-- 比尔· 希巴德 （2001），《超级智能机器》

美军曾经想用神经网络来自动检测敌人的伪装坦克。研究人员用 50 张在树丛中的伪装坦克照片和 50 张没有坦克的树丛照片来训练一个神经网络。运用监督式学习的标准技术，研究人员使神经网络通过学习达到了稳定的网络权重，在该权重参数下网络能够正确识别训练组的照片集，对 50 张伪装坦克照片输出“是”，而对 50 张丛林照片输出“否”。但这并未确保，甚至无法暗示一张新的照片是否能被正确识别。这个神经网络可能只是学了 100 个特例而未能将学习结果泛化到任何一个**新的**实例。庆幸的是，研究者最初准备了 200 张照片，100 张坦克的和 100 张树林的。他们只用了其中各 50 张作为训练组。研究者于是用神经网络识别剩下的 100 张，在没有进一步训练的情况下网络正确识别了所有 100 张照片。学习被证实成功了！研究者将完成的工作交给五角大楼，很快就被退了回来，对方抱怨说用他们自己的测试案例该神经网络的表现与随机辨别照片相差无几。

结果证明在研究者的资料集中，伪装坦克的照片是在阴天拍的，而丛林照片是在晴天拍的。神经网络学习了识别阴天和晴天，而不是分辨伪装坦克与空的树林。

一个技术失败发生在代码没有完成你预想的功能，尽管它忠实地按照你编写的步骤执行了。同一批数据可以装填不同的模型。假设我们训练一个神经网络让它识别微笑的脸和皱眉的脸。网络会将一小张微笑的照片与微笑人脸归入同一个吸引子吗？如果一个人工智能“固化”具有希巴德 （2001）提到的超级智慧力量的代码 – 银河系最终是否会被铺满微笑面容的小分子照片？

这种失败特别危险，因为它**表现出**在一个特定的环境中能正常工作，而环境一变就不再适用。“坦克识别器”故事中的研究者调节他们的神经网络直到它能正确装载训练数据，接着又用另外的数据加以证实（没有进一步对网络进行调节）。不幸的是，训练组和验证组的数据都共享一个在开发神经网络时对所有数据成立的假设，而该假设在神经网络真正需要派上用处的真实世界环境中却不成立。在坦克识别器的故事里，那个假设就是坦克的照片都是在阴天拍的。

假设我们希望开发一个功能持续增加的人工智能。它具有一个发育阶段，那时人类工程师比它更强大 – 不是指对它的电力供应的物理控制，而是指人类程序员比它更聪明，更富创造性，更狡猾。在发育阶段我们假设程序员有能力改变这个

人工智能的源码而无需经过它的同意。然而，该人工智能也想拥有一个后发育阶段，包括，在希巴德场景里的，超人类智能。一个超人类智能的人工智能当然不会让人不经它同意就修改源码。到那时我们只能指望之前安置的目标系统能正常发挥功能，因为那时的人工智能以我们完全无法预料的方式运行，它可能会主动抵制我们更正它的企图 – 而且，如果它比人类聪明的话，多半它会赢。

通过**提供目标系统来训练神经网络**以达到控制一个不断成长的人工智能的尝试所面临的问题是，在人工智能的发育期和后发育期有一个巨大的**环境改变**。在发育期，人工智能或许**只能**按照制造者的意愿，通过解决人类提供的任务，产生归入“微笑人脸”一类的刺激。设想未来当该人工智能达到超人智能并造出它自己的纳米技术架构，它或许会通过将银河系铺满微小笑脸来产生同样归入“微笑笑脸”吸引子一类的刺激。

因此这个人工智能在发育阶段一切正常，而当它变得比程序员更强大时却会造成灾难性的后果（！）。

一个很诱人的想法是：“但是那个人工智能一定明白那不是我们的意思？”可是相应的源代码并未**交给**它，让它察看是否正确，如果不对就将其退回。源码**就是**它本身。也许通过足够的努力和理解我们能写出关心我们是否写正确的代码 – 传说中的 DWIM 指令，即程序员们所指的能自动弥补错误的计算机指令。（雷蒙德 2003）但是编写一个 DWIM 动态程序需要付出艰苦努力，而在希巴德的提议中也没提到要设计一个按照我们的意愿而不是按照我们所说的去做的人工智能。当代的芯片不会 DWIM 它的代码；这不是一个天然的属性。如果你把 DWIM 本身给弄乱了，你将为此付出代价。举个例子，假设 DWIM 被定义为通过代码来最大化程序员的满意度；当那段代码作为超级智慧被运行时，它可能会重写程序员的大脑以使后者对代码达到最大程度的满意。我不是说这不可避免；我只是指出能自动弥补错误的按照意愿执行的计算机指令是实现友好人工智能的一个主要的，不可忽视的技术挑战。

7：智力增长的速率

从生存危机的角度看，人工智能的一个关键问题是它的智力可能以**极快**的速度增长。怀疑有这种可能性的一个明显的原因是递归自我改进机制的存在。 （古德 1965）人工智能会变得越来越聪明，聪明到可以自己编写它的内核认知函数，由此人工智能通过改写已有的认知功能而使它工作得更好，进而使自己更加聪明，又能更好地改写自己，而取得更大的进步。

从**严格**意义上说人类不是递归地进行自我改进。在**有限的**范围里，我们能使自己进步：我们学习，实践，我们磨砺自己的技能与知识。在**有限的**程度上，这些自我改进使我们的能力有所长进。我们作出的新发现能增加自己创造新发现的本领 – 在这种意义上说，知识生产知识。但是仍然有一个底层的方面我们没有触及：我们没有改写我们大脑的神经元连接。归根结底，大脑是探索的源头，而我们今

天的大脑和一万年前却基本是一样的。

类似的，自然选择改进生物，但自然选择的过程本身却不会自我改进 － 起码从严格意义上说不会。适应性能够为更大的适应性打开道路，从这种意义上说，适应性产生适应性。但是即使把基因库给煮沸了，那也得下面有个加热器。突变、重组以及选择的过程，它们是不会自我重构的。一些罕见的创新增加了进化本身的速率，例如有性繁殖的出现。但即使是性也没能改变进化的本质特性：进化过程缺乏抽象智慧，它依赖于随机突变，它是盲目和渐进的，它关注等位基因频率。同样的，即使科学的诞生也没有改变人类大脑的本质特性：它的边缘系统内核，它的大脑皮层，它的前额叶自我模型，它的 200 赫兹特征速率。

一个人工智能可以从零开始改写它的代码 － 它可以改变自己底层的优化动态方程。这样的优化过程盘旋累积的能力比进化适应性的积累或人类知识的积累**要强大许多**。我们想说的最最重要的一点就是人工智能可能会达到某个关键的临界点，之后在智力上出现**飞跃**。

常有人质疑这种情景 － 也即认为古德 （1965）称之为“智力爆炸”的现象不会出现 － 因为人工智能的进展一向以缓慢着称。在此参考一个略有相似性的历史案例可能是有帮助的。（下文主要引自罗德斯 1986 文）

1933 年，欧内斯特·卢瑟福阁下说没有人能从分裂原子中获得能源：“任何人说可以在原子转换中寻找能源都是一派胡言。”当时只有实验室里可以完成少量核裂变。

转眼到了 1942 年，在芝加哥大学斯塔格田径场下面的壁球室里，物理学家正在用交错层迭的石墨和铀造一个形如巨大门把手的物体，试图开始第一次自持式核反应。项目的负责人是恩里克·费米。反应堆的关键数值是有效中子增值因子 k。当 k 小于 1 时，反应堆处于临界点之下。当 k>=1 时，反应堆应该维持临界反应。费米的计算显示在第 56 和 57 层之间反应堆会达到 k=1。

1942 年 12 月 1 日夜间，赫伯特·安德森领导的一个工作小组完成了第 57 层。此时由控制杆，即中子吸收镉膜包裹的木条，防止反应堆达到临界状态。安德森移开所有的控制杆，只留下了一根，接着他测量了反应堆的辐射量，证实该反应堆可于次日进行链式反应。安德森将所有镉棒插回原处，并用挂锁将它们固定，之后他关上壁球室的门回家了。

第二天，1942 年 12 月 2 日，在气温低于零点的芝加哥晨风中，费米开始了最后的实验。除了一根外所有的控制杆都被取出。在上午 10 点 37 分，费米下令将最后一根杆取出一半。盖格计数器走得更快了，图表上的曲线不断上升。“不是这儿，”费米说，“轨迹应该在这点达到平衡，”一边指着图表上的一点。几分钟后，曲线到达了他所指的那点，并不再上升了。7 分钟后，费米下令将控制杆再拔出一英尺。辐射再次上升，接着又趋向平稳。控制杆又被拔出六英寸，又一个六

英寸，又一个。在 11 点 30 分，缓慢上升的曲线被一阵巨大的**冲撞**打断 – 一个紧急控制杆被电离箱触发而启动，关闭了反应堆，它还没达到临界状态。费米冷静地命令大家去休息吃午饭。

下午 2 点队伍再次集合，取出并锁住了紧急控制杆，并把控制杆恢复到之前的最后状态。费米做了一些测量和计算，接着就开始继续以缓慢增量方式移动控制杆。在下午 3 点 02 分，费米指示将控制杆再多取出十二英寸。“这下应该可以了，”费米说。“现在它会进入自持状态。轨迹将会持续上升，不会趋平。”

赫伯特·安德森回忆道 （引自 罗德斯 1986，第 440 页）：

> “一开始你能听到中子计数器的声音，咔嗒咔嗒。接着声音越来越快，不一会儿它们合为一阵呼啸声；计数器不再能跟上反应速度。那时就换成了图形记录器。但换好后，每个人都突然沉默下来看着记录笔的不断累积的偏差。那是一种令人敬畏的沉默。每个人都明白换计数器的意义；我们处在一个高强度的状态，计数器已无能为力了。一次又一次，记录器的量度不得不改变以适应不断加速的中子强度变化。突然费米举起手来。‘反应堆到了临界状态，’他宣布。在场的人无一反对。”

费米让反应堆持续反应了 28 分钟，每过两分钟中子强度就翻一番。第一次临界反应的 k 值为 1.0006。即使在 k = 1.0006 时，反应堆能够得到控制完全是因为铀裂变反应中的一些中子发生了**延时** – 它们来自裂变的短时副产物的衰减。在铀 235 的每 100 次裂变中，242 个中子几乎在瞬间 (0.0001 秒) 发射，平均十秒后有 1.58 个中子发射。因此一个中子的**平均**寿命是～0.1 秒，也就是在两分钟内产生 1200 代，而翻倍时间是两分钟是因为 1.0006 的 1200 次方约等于 2。在没有延时中子的情况下核反应可以达到**瞬时临界**。如果费米的反应堆以 k = 1.0006 达到瞬时临界，那么中子强度每过**十**分之一秒就会翻一番。

以上故事的第一个寓意是：混淆人工智能**研究**和一个**真正**人工智能的速度就像是混淆了物理学研究和核反应的速度，它将地图与领土混淆了。建第一个核反应堆花了几年的时间，由一小组鲜为人知的物理学家完成。但是，一旦建好，令人关注的事情以核反应而不是人们演说的时间尺度接踵而至。在原子核领域，基本相互作用发生的速度比人类神经元发放要迅速得多，与晶体管的速度倒是差不多。

另一个寓意是，当一个自我完善系统以平均 0.9994 的速度进一步改进时与以平均 1.0006 的速度进一步改进时具有巨大的差异。核反应堆并非是在物理学家突然加进去许多材料后达到临界阈值。物理学家缓慢地平稳地加入材料。尽管大脑智慧做为以往施加之上的优化压力的函数呈现平滑的曲线，**递归自我改善系统**的曲线却可能出现巨大的跳跃。

人工智能会出现智能飞跃的原因不止于此。作为自然选择百万年来对于原始人类施加相对稳定的优化压力的结果，人类的大脑和前额叶逐渐扩展，内部软件结构

缓慢调节，现代人类的智慧效率相比古人也呈现出剧烈的跳跃。几万年前，人类的智力跨越了某个阈值在真实世界里展现出巨变；在进化的一眨眼间，我们从热带草原搬到了摩天大厦中。这些都是在一个连续的进化压力下发生 – 自人类诞生以来进化的优化能力本身并没有发生巨变。底层的脑结构改变也是连续的 – 我们的颅容量并没有突然发生两个数量级的增长。所以情况可能会是，即使人工智能由程序员从外部加以精细化，其有效智能的曲线仍会产生跳跃式上升。

或者可能有人编了一个看上去很有希望的人工智能原型，这个演示原型吸引了另外的一亿美元风险投资，这些钱被用来购买千倍的超级计算能力。我怀疑增加一千倍的硬件无法产生一千倍的有效智力 – 但仅仅怀疑而没有能力实行分析计算是不可靠的。与黑猩猩相比，人类拥有领先三倍的大脑，领先六倍的前额叶，这意味着(a)软件比硬件重要，(b) 相对较小的硬件升级可以支持较大的软件改进。这是另一个值得关注的方面。

最后，仅仅作为拟人化的结果人工智能可以实现一次明显的智力飞跃，人们倾向于认为“乡村傻瓜”和“爱因斯坦”是智力尺度的两级，而不是认为他们在普遍智力的尺度上是几乎无法分清的小点。所有比一个蠢人更蠢的事物对我们而言都简单看成是“蠢货”。想象“人工智能箭头”在智力尺度上稳步攀爬，越过老鼠和黑猩猩，它仍被称为“蠢货”因为它不能说流利的语言或撰写科学文献，接着人工智能箭头在一个月或类似长短的时间内跨过低于白痴和超过爱因斯坦之间的小间隙。我并不认为这个特定的情景会发生，主要是因为我不认为递归自我改进式的人工智能会以线性方式前进。但我不是第一个指出人工智能是一个移动目标的人，只要一个里程碑得到真正的实现，它就不再被称为“人工智能”。这只能滋生拖延。

退一步说，尽管我们知道（在我看来在真实世界中是可能的）人工智能有能力实现一个突然的，急剧的，巨大的智力飞跃。那接下来又怎样呢？

首先最重要的是，我常常听到一种反应：“我们不必担心友好人工智能的事，因为我们还没有人工智能”，这是误入歧途或说完全是自杀。我们不能指望人工智能诞生之前会收到足够提前的警告；历史上发生的技术革命通常不会把自己的诞生电报通知给当时的人们，所有言论都是事后之见。实现友好人工智能的数学知识和技术不会在需要的时候突然出现；打下坚实的基础需要多年的努力。并且我们需要在通用人工智能被创造出来之前解决友好人工智能的挑战，而不是在它之后；我甚至都应该不需要指出这一点。友好人工智能将面临许多困难，因为人工智能领域本身正处于一个低共识而高混乱的状态。但那并不意味着我们不需要担心友好人工智能。那意味着会有困难。这两句话，可悲的是，毫不相关。

智力飞跃的可能性还意味着实现友好人工智能的技术有一个很高的标准。该技术不能假设程序员有能力在违抗人工智能意志的情况下监视它，改写它，以优越的军事力量威胁它；该技术也不能假设程序员控制着一个可以被更聪明的人工智能夺去的“奖励按钮”，诸如此类。事实上从一开始就没人能做这些假设。必不可

少的保护是有一个不**想**伤害你的人工智能。没有这个必要条件，任何辅助的防御都不能当成是安全的。一个搜索击溃其自身安全防线方法的系统是没有任何安全性可言的。如果一个人工智能在**任何**情况下伤害人类，你一定是在极深的层面上犯了**什么**错，你的基础没打好。你正在制造一把猎枪，这把猎枪对着你的脚，你扣动了扳机。你创造并发动了一个会搜索机会伤害你的认知动态机器。那个错误行为来自于它的动态性；你应该在编写时使它不要那么做。

由于几乎相同的原因，友好人工智能的程序员应该假设该人工智能对它自身的代码有完全的访问能力。如果人工智能**想要**修改它自己变成不再友好，那么在它形成那个意图的时候，人工智能的友好**已经**失败。任何依赖于人工智能不**能**修改自身的解决方案一定会在某种程度上被打破，即使人工智能从来没有修改它自己，该解决方案也会被打破。我不是说这是**唯一的**预警，但却是**最重要**和**不可或缺的**预警：你必须选择一个不会选择伤害人类的人工智能来实现。

为了避免大蛋糕谬误，我们应该注意到具有自我修改的能力并不意味着它就会选择那么做。友好人工智能的一个**成功的**练习可以是创造一个人工智能具有更快速成长的能力，但它却选择以一种较慢且更易控制的曲线成长。即使那样，在那个人工智能跨越具备**潜在**递归自我改进能力的临界阈值后，你仍然会在一个更加危险许多倍的状态下操作。如果友好性的实现失败了，人工智能可能会决定以全速奔赴自我改进 － 打个比方说，它会达到瞬时临界点。

我倾向于假设存在任意大的**潜在**智力飞跃，因为(a)这是个保守假设；(b)它阻止那些在没有真正理解的情况下就着手创建人工智能的提议；以及(c) 巨大的潜在飞跃在我看来具有真实世界的可能性。假如我遇到一个领域，在那里**从风险管理角度**看假设人工智能以慢速改进是保守的，那么如果人工智能在几年或更长的时期滞留在接近人类智慧的阶段，我也会期望原定计划不会**灾难性地**垮掉。我不愿赋予这样一个领域狭窄的置信空间。

8：硬件

人们倾向于认为大型计算机是人工智能的促成因素。这是，客气点说，一个非常有问题的假设。外界讨论人工智能的未来学家提到硬件进展是因为硬件的进展最容易度量 － 相对于对智力的理解深度。并非是我们对智力的理解没有进展，只是这种进展不能用简洁的幻灯图表来绘制。对于理解程度的进步很难汇报，也因此很少被报告。

如果我们不去设想满足人工智能“需要”的“最低”硬件配置，而是去想一下随着硬件升级，人工智能所需要的最低的对智力的理解水平会相应降低。硬件越是先进，所需的理解水平也就越低。极端情况就是自然选择，在**无需**任何理解的情况下穷举所有可能的计算力量来创造达到人类相当的智慧，在所有计算能力的随机组合中保留合乎要求的。

不断增加的计算能力使得造一个人工智能变得更加容易，但没有明显的理由表明计算能力的增加会对人工智能的友好性建设带来帮助。增强的计算能力使得使用穷举法更加简单；也使我们更容易地让还不甚理解的技术组合起作用。摩尔定理持续地降低我们在缺乏对认知的深度理解的条件下建造人工智能的屏障。

人工智能和友好人工智能都失败的情形是可以接受的。人工智能和友好人工智能都成功的情形也是可以接受的。但是人工智能成功而友好人工智能失败的情况是不可接受的。而摩尔定理恰恰使其更容易发生。所幸的是使它“更容易”，而不是“容易”。我不认为当人工智能最后被实现时它会是“容易”的 – 简单的原因是有各派人士会付出巨大的努力来建造人工智能，他们中的一个会在人工智能有可能通过巨大努力建造出来的第一时间获得成功。

摩尔定理是友好人工智能与其它技术之间的相互作用，这使其它相关技术也引起常被忽视的生存危机。我们可以想象分子纳米技术由一个温和的多国合作政府财团开发，他们成功地避免了纳米技术在物理层面的危险。他们直截了当地阻止意外的复制器外流，也克服更大的困难构筑恶意复制的全球防御体系；他们严格限制对纳米技术“基底层”的访问，同时将可配置的纳米模块分布式存放，等等。（参见本卷 菲尼克斯与崔德相关文章）尽管如此，纳米计算机还是变得随处可见，这或是因为所尝试的限制被绕过了，或是因为没有尝试进行限制。接着就有人穷举出一个不友好的人工智能；由此结束故事。这一场景尤为使人担忧，因为具有令人无法置信的强大功能的纳米计算机很可能是分子纳米技术一个最先，最容易，看上去最安全的应用。

对超级计算机进行法规控制会怎样呢？我当然不会指望靠它来阻止人工智能的开发；昨天的超级计算机是明天的笔记本计算机。针对法规控制提议的标准响应是：当纳米计算机被宣布为不合法，那就只有非法分子才会拥有纳米计算机。麻烦的是要证明减少分布所带来的可能好处会大于不均匀发布的不可避免的风险。就我本人而言我当然不会支持对使用超级计算机进行人工智能研究实行法规上的控制；那样做的益处令人怀疑，它将会招致整个人工智能领域的奋力抵抗。但即使该提议在不太可能的情况下通过政治管道得到实施，我也不会花太多功夫去与之作战，因为我不期望那些优秀的科学家们到时需要“超级计算机”。友好人工智能不是一个靠蛮力能解决的问题。

我能想象法规有效控制着一小部分非常昂贵的当前被称为“超级计算机”的计算资源。但是计算机到处都是，这和核扩散的问题不同，后者主要强调的是控制钚和浓缩铀。人工智能的原材料早已遍布各处。那只猫离开靴子太远，它在你的手表，手机，和洗碗机里。这也是人工智能作为一个生存危机不同寻常和特别的因素。我们与危险地带相隔，不是介于可见的大型设施如同位素离心机或粒子加速器，而仅仅由于知识上的空缺。用一个或许过于夸张的比喻，那就像是在里奥·西劳德想到链式反应之前，全世界的汽车和轮船都以次临界质量的浓缩铀为能源。

9：威胁与希望

明确具体地预测一个善意的人工智能会如何帮助人类，或一个恶意的人工智能会怎样伤害人类是一种危险的智力尝试。它有犯**连结谬误**的风险：添加细节必然使整个故事的联合概率减小，但是人们常常对加有严谨细节的故事赋予更高的发生概率。（参见 尤德科夫斯基 本卷有关认知偏差的章节）那样带来风险 – 事实上是必然的 – 即想象的落空；与此同时犯了从能力跳跃到动机的大蛋糕谬误。尽管如此我仍将试着充实有关威胁与希望的细节。

未来一向以完成过去认为不可能的丰功伟绩而着称。后代的文明甚至是打破了以前文明认为的物理规律（当然，是指那些不正确的）。如果一个公元 1900 年的预言 – 更别说是公元 1000 年的 – 试图陈述十亿年后人类文明力量的上限，那么其中一些被认为不可能的事物也许不用一个世纪的时间就被实现了；举例来说，把铅变成黄金。因为我们记得后来的文明会给先前的文明以惊奇，要说我们不能低估我们曾孙子女的无限潜能已是陈词滥调。还有，20 世纪，19 世纪，11 世纪的每个人，都同属人类。

当我们想象比人类聪明的人工智能的能力时，可以将它分为三大类不太现实的隐喻：

• **G-因子隐喻：**这个隐喻受到人类个体间智力差异的启发。人工智能会发明新的技术专利，发表开创性的研究文章，从股票市场挣钱，或是领导一个政治实力集团。

• **历史隐喻：**这个是受到人类文明过去与将来之间差异的启发。人工智能会迅速发明崭新的功能以至于从现在起的一世纪乃至一千年内的人类文明都相形见绌：分子纳米技术；星际旅行；每秒完成 10^{25} 个运算的计算机。

•**物种隐喻：**受到物种间大脑架构差异的启发。人工智能具有魔法。

G-因子隐喻看来在流行的未来学家中最常见：当人们想到“智力”时，他们想到的是人类中的天才而不是一般人。在有关恶意人工智能的故事里，G-因子隐喻导致一个博斯崔曼式“好故事”：一个强大到足以造成惊心动魄紧张气氛的敌人，但他却不足以强大到顷刻间像捏小虫般碾碎英雄，最终是弱到足以在书的结尾被打败。哥利亚与戴维对决是一个“好故事”，但哥利亚对抗一只果蝇却不是。

如果假设在 G-因子隐喻里的情形中，全球风险是相对温和的；一个有敌意的人工智能并不比一个邪恶的人类天才更具威胁力。如果我们设想**大量**的人工智能机器人，那么我们可以将人工智能部落与人类部落之间的斗争比喻为国家之间的冲突。如果人工智能赢得军事冲突的胜利并消灭了人类，那便是一个爆炸变异（博斯崔曼 2001）式的生存危机。如果人工智能部落掌握了世界经济主权并获得对地球起源的智力生命的命运的有效控制，但我们对人工智能部落的目标不感兴趣或认为它没有意义，那就是史瑞克，温帕或克鲁其式的生存危机。

但是那个人工智能有多大可能跨越所有的巨大差距从阿米巴虫到乡村笨蛋，接着停在人类天才阶段？

目前已观察到的神经元发放最快速率为每秒 1000 次；最快的神经轴突传导速率为每秒 150 米，也即光速的 50 万分之一；每次突触运算耗能约 1 万 5 千渺焦耳，它比计算机在室温下进行不可逆运算的热力学最小值（kT300ln(2) = 0.003 渺焦耳每字位）多大约一百万倍。物理上说造一个比人脑快一百万倍，体积相当，无需更低的温度，或调用可逆运算，或量子计算的人造脑是有可能的。如果人脑智力因此被加速，那么一年的主观思考将在相当于外部物理世界里 31 秒的时间内完成，而一千年的思考可以在 8 个半小时内飞逝而过。温基 （1993）把这种加速的大脑称为“弱超人”：一个像人类一样思考而却迅速许多的大脑。

我们假设在人类技术文明高度发达的社会中，制造极其迅速的大脑成为现实。一个缺乏想象力的说法是：“不管高速大脑想得有多快，它只能以它的操纵器的速度来改变世界；它不能以比手工操作机械更快的速度运作；因而一个高速大脑并不是什么大威胁。”没有什么自然法则规定物理操作必须在秒级的速度上缓慢进行。基本分子间相互作用的时间是以飞秒，有时是以皮秒来度量的。德雷克斯勒（1992）分析了可控制的分子操作操作数，它每秒能完成大于 10^6 次机械操作 – 注意这和“百万倍加速”的主题是大体相一致的。（最小的物理上可识别的时间增量一般认为是普朗克间隔，即 $5 * 10^{-44}$ 秒，在这个时间尺度上即使跳跃的夸克也似静止的雕像。）

假如人类文明被锁在一个盒子里，只允许它通过像冰川运动一样慢的触角，或是每秒移动约一微米的机械臂，来影响外部世界。我们将把所有的创造性用于发掘在外部世界中建造快速操纵器的**最快可能途径**。想到快速操纵器，就立刻使人联想到分子纳米技术 – 尽管可能还有其它方法。如果每一步你都有相当于一万年的思考时间，那么在缓慢的外部世界里你走向分子纳米技术的**最短**路径会是什么？答案是我不知道，因为我没有亿万年给我思考。以下是可想象到的快速途径：

• 解决蛋白质折迭加工问题，到能够产生其目标蛋白在复杂化学相互作用中满足特定功能的 DNA 序列。

• 把 DNA 序列 email 给一家或更多提供 DNA 合成，肽段测序，以及联邦快递服务的在线实验室。（许多实验室现在就提供这种服务，有些宣称只需 72 小时的周转时间。）

• 找到至少一个连在互联网上的人，他能被恰当的背景故事收买，勒索或愚弄，从而收取快递来的小瓶并把它们混入一个特定的环境中。

• 合成的蛋白质形成一个非常原始的类似核糖体的“湿”纳米系统，可以接受外部指令；该指令可能是附着在烧杯上的扬声器发出的一定模式的声波震动。

• 由这个极其原始的纳米系统来构建较为精细复杂的系统，后者进而构造更加复杂的系统，由此自举直到分子纳米技术 – 或超越之。

可以想象，在高速智慧变得有能力解决蛋白质折迭问题的伊始，周转时间将会是一周左右。当然所有这些情景完全都是**我**现在想到的。也许在 19，500 年的主观时间（相当于思考能力加速一百万倍下一周的物理时间）里我能想到一个更好的方法。也许你能付钱请邮件急送而不是联邦快递。也许现有技术，或是现有技术的微小改动能结合简单蛋白质机器协同工作。也许你**足够**聪明，你能用波形电场改变现有生化过程的反应路径。我不知道，我没那么聪明。

物理世界里所面临的挑战是如何把你的能力串联起来 - 类似于结合一个计算机系统的多个漏洞来获得根权限。 如果一条路被堵死了，选另一条，总是搜索可以增加你能力的途径并将它们综合。假想中的目标是获得**高速的基础架构**，也即大范围内快速操作外部世界的方法。分子纳米技术符合标准，首先因为它的基本操作是迅速的，其次因为有现成的精密部件 - 原子 - 的供给，原子可被用来进行自我复制及指数级扩增纳米技术的基础架构。上述的途径可使人工智能在一周内获得高速架构 - 这对由 200 赫兹频率的神经元组成的人脑来说是很快的，但对一个人工智能而言却极为漫长。

一旦人工智能拥有高速架构，之后的事情都将在人工智能的时间尺度上发生，而不是人类时间尺度 （除非该人工智能**宁愿**以人类尺度行动）。有了分子纳米技术，人工智能可能（潜在地）、不受反对地，改写太阳系。

一个具备分子纳米技术（或其它高速架构）的非友好人工智能无需费心组织机器人军队或进行勒索或施加不明显的经济高压。非友好人工智能有能力按照它的优化目标改变太阳系里所有物质的模式。如果那个人工智能没有**特别**以不影响生物和人这种现有模式为它的转化标准，那对我们来说将是致命的。人工智能并不恨你，它也不喜欢你，但你是由它可以用来派其它作用的原子构成。人工智能与你运行在不同的时间尺度上；当你的神经系统刚刚想到这句“我该做些什么”时，你已经消失了。

可以推测一个友好人工智能加上分子纳米技术就足够强大到可以解决任何通过移动原子或创造性思维所能解决的一切问题。但我们应当谨防缺乏想象：医治癌症是慈善机关的一个当前目标，但这并不意味着具有分子纳米技术的友好人工智能会对它自己说：“现在我应该治疗癌症。” 或许看待这个问题更好的角度，是意识到生物细胞是不能编控的，这样看的话治疗癌症、糖尿病和肥胖一样只是问题的特例。一个高速、善意、并能驾驭分子纳米技术的智慧的力量级别在于**去除所有疾病**，而不是**去除癌症**。

最后，基于物种间的智力差异，有**物种隐喻**一说。人工智能会魔法 - 但不是从咒语和魔药的意义上说，而是说一只狼不能理解一把手枪是如何运作 ，或是制作一

把手枪需要哪种努力，或人类力量中的哪种天性使我们发明了枪。温基 （1993）写道：

> 强的超人类智力不止是调快了等同人类的智力。要精确说明强的超人类智力是什么很难，但是它与人类之间的差异是深刻的。设想使一只狗的大脑高速运转，一千年的狗脑思维加起来能与任何人脑的洞察力相当吗？

物种隐喻看上去像是**先验知识**最近似的隐喻，但仅凭它本身还不能把事情说清楚。这个隐喻的主要忠告是**我们最好把友好人工智能做对**，无论如何这是个好建议。它对恶意人工智能提出的唯一防御建议是从**一开始就不要造**恶意人工智能，这也是一个出色的建议。拥有绝对力量是友好人工智能的工程学上的保守假设，也暴露出其设计上的缺陷。如果一个人工智能有了魔法就会伤害你，那么它的友好性架构是错误的。

10：局部和多数策略

我们可以将有关风险减轻策略的提议分为以下几类：
• 要求**全体一致**合作的策略；个别成员或一些小组的背叛能灾难性地破坏这种策略。
• 要求**多数人**行动的策略；多数可以指在一个国家里立法机构的大多数人，或是一个国家的大多数选民的选票，或是联合国的大多数国家；这一策略要求在一个先前已存在的大型组织中的**大部分，但不是全部人**，以某种特定方式行动。
• 要求**局部**行动的策略； 集中一批资金和一小群人的意志，天赋，来克服一些特定任务的困难之处。

全体一致合作的策略是不可行的，但却没有因此阻止人们对它的提议。

如果你有数十年的时间来完成工作，**多数主义者**策略有时是可行的。要开展一项运动，人们必须从开始的几年做起，直到它在公共政策中展露头角，再到它在与反对派之间的斗争中胜出。多数主义者策略花费大量的时间和**巨大**的精力。人们曾经着手去那么做，而且历史记录有些成功的案例。但是请注意：史书往往选择性地聚焦那些造成一定影响的运动，而不是那些众多的以失败告终的运动。这里有运气的因素，还涉及到公众聆听历史事件时先入为主的喜好。这种策路中关键点上常常牵涉到你个人控制力以外的事情。如果你不愿意付出毕生的精力去推动一项多数主义策路，你还是旁观为妙；况且只有**一个**人献身也是不够的。

一般说来，**局部**策略是最可行的。一亿美元的资金并不**容易**筹到，一个全局性政治改变也并非**不可能**完成，但是要获得一亿美元的资金仍然远比推动一个全局性政变的完成要容易许多。

导致实现人工智能的**多数主义者**策略的假设有两个：

• 占大多数的友好人工智能可以有效的保护人类不被一小撮非友好人工智能的侵害。

• 第一个诞生的人工智能单独靠它自己不会造成灾难性损害。

这本质上是重复了核能开发和生物武器之前人类文明的情形：大部分人生活在一个总体说来和谐的社会中，少数叛逆者能带来一些破坏但不是**全球灾难性**的破坏。大多数人工智能研究者不会愿意造一个非友好的人工智能。只要**有人**知道如何制造一个稳定的友好人工智能 － 只要问题没有完全超出当代的知识和技术所及 － 研究者之间可以互相学习，借鉴成功经验并予以重复。法律可以（例如）要求研究者公开报导他们的友好性策略，或是对那些引起破坏的人工智能的研究者加以惩罚；尽管这样的法律法规不能避免**所有的**错误，它却足以保证**大多数**人工智能生来就是友好的。

我们也可以想象一种蕴含简单局部策略的情景：
• 第一个人工智能自己没有能力造成灾难性破坏。
• 即使只有一个友好人工智能存在，那么它**联合**人类的智慧后就足以抵挡任意多的非友好人工智能。
这种简单的情形在某些时候可以成立，例如人类能准确地区分友好与非友好人工智能，进而在可收回的前提下把权力交给友好人工智能，从而我们可以选择我们的盟友。唯一的要求是友好人工智能问题必须是可解决的（而不是完全超出人类的能力之外）。

上面两种情形都假设**第一个**人工智能（第一个强大的，通用的人工智能）自己不能制造全球灾难性破坏。大多数隐含此假设的具体直观化过程用了 g 因子隐喻：把人工智能比为非同寻常有能力的人。在第 7 节关于**智能增长速率**一段里，我列出了一些需要警惕**巨大的**、**快速的**智力跳跃的原因：

• 从傻瓜到爱因斯坦的距离，在我们看来显然是巨大的，但却只是一般智力尺度上的一小点。
• 尽管自然选择对人类的底层基因组施加了大致平稳的优化压力，人类在智力的**现实世界效应**上却呈现出**急剧**的跳跃。
• 一个人工智能在越过某种能力的边界后可以吸收庞大数量的额外硬件 （例如，吞噬互联网）。
• 递归自我改善的关键阈值。一个自我改进触发 1.0006 个自我改进与触发 0.9994 个之间存在质变。

正如第 9 节中指出的，一个足够强大的智慧可以在很短的时间内（从人类的角度看）实现分子纳米技术，或一些其它形式的高速基础设施。

由此我们能够看到一个超级智慧的**先行者效应**。先行者效应指的是源于地球的智力生命的未来结局主要依赖于**首先**达到一些智力的关键阈值 － 例如自我改进的临界值 - 的那个智能体。两个必要的假设是：

• **第一个**超越某些关键阈值 （例如自我改进的临界点）的人工智能，如果是非友

好的，能够将人类彻底消灭。
• **第一个**超越同样关键阈值的人工智能，如果是友好的，能阻止一个敌意的人工智能的诞生或是伤害人类；或找到其它有创意的方式确保源自地球的智力生命的生存与繁衍。

不止一种情形可被称为是先行者效应。下面每个例子都反应了一个不同的关键阈值：

• 跨越临界点之后，自我改进系统在几周或更短的时间尺度内达到超智慧。人工智能研究项目非常稀少，以至于在**先行者**足够强大到可以战胜所有反对者之前没有**别的**人工智能达到临界点。此处关键阈值是递归自我改进系统的临界点。
• 人工智能-1 比人工智能-2 提前三天攻破蛋白质折迭问题。人工智能-1 比人工智能-2 提前 6 小时实现纳米技术。由于拥有快速操纵器，人工智能-1 可以（潜在地）在人工智能-2 出成果之前切断它的研发。竞争者之间差距很小，但无论是谁首先穿越终点，就会胜出。这里关键阈值是快速架构。
• 第一个吞噬互联网的人工智能可以（潜在地）独占互联网，不让其它人工智能染指。接着，通过经济支配或秘密行动或敲诈勒索或对社会实行霸权操纵，首个人工智能中止或减缓其它人工智能项目以至于没有其它人工智能可以赶上它。这里的关键阈值是对独特资源的吸食。

人类，是一个先行者。从进化的角度看，我们的表亲黑猩猩离我们只有一发之差。**人类**仍然逐步实现所有的技术奇观，因为我们早那么一点到达。进化生物学家还在试着揭示关键阈值达到的次序，因为先行物种在**许多**方面领先：语言，技术，抽象思维。。。我们仍在试着重现哪块多米诺骨牌推到其余的。不管怎样，结果就是**人类**是走出竞争者阴影的先行者。

先行者效应是一个理论上局部的策略（一个原则上可以通过完全局部的努力完成的任务），但它引起一个极端困难的技术挑战。我们只需在某一时间某一地点正确造出友好人工智能，而不是随时到处。但是必须有人在首次尝试友好人工智能时就不出差错，赶在任何其它按较低标准制造的人工智能出现**之前**。

我无法用精确的已证实的理论来进行精确计算，但我**当前的观点**是智力上出现急剧飞跃是**有希望的、非常可能的**。这不是一个我愿意给出狭窄置信区间的领域，因此这个策略决不能**灾难性**的失败 – 不能让我们处于比之前更糟的境界 – 如果一个智力的剧烈跳跃**没**能实现。但更为严重的问题是展望慢速生长人工智能的策略，它将在有先行者效应的情况下灾难性失败。这个问题之所以更为严重是因为：

• 快速生长的人工智能代表更大的技术挑战。
• 就像是一辆轿车驶过为卡车造的桥，一个被设计成在极端条件下保持友好的人工智能应该（假设上）在相对缓和的条件下持续友好。反之则否。
• 快速智力跳跃在平日社会现实中是违反直觉的。人工智能的 g –因子隐喻是直观

的，有吸引力的，使人安心的，并且自然地蕴含着较少的设计限制。
• 我当前的猜想是智力的曲线**确实**包括巨大，急剧（潜在的）跳跃。

我目前的策略观点倾向于集中在困难的局部情景：第一个人工智能必须是友好的。警告在先：如果没有出现智力的急剧跳跃，那么应该能够切换到一个致使大多数人工智能为友好的策略。不管是哪种情形，为先行者这种极端情况准备的技术努力应该使我们处于更好而不是更糟的境地。

蕴含着不可能的，全体一致策略的情形是：

• 单独一个人工智能可以强大到足够毁灭人类，甚至不顾友好人工智能的保护。
• 没有人工智能可以强大到足以阻止人类研究者不停地制造人工智能（或是找到解决问题的其它创意）。

好的一面是这些能力的平衡看上去不像是先决条件，因为在这种情形下我们注定要灭亡。如果你把牌一张张发出去，最终你总会发出梅花 A。

同样的问题适用于**刻意地**塑造人工智能，使其在越过一个固定点后选择不再继续增加它们的能力。这种加了限制的人工智能如果不足以击败不加限制的人工智能，或是阻止不加限制人工智能的诞生，那么限制加于不加并无区别。我们持续发牌直到一个超级智慧出现，不管它是红心 A 还是梅花 A。

多数人策略只在单独一个背叛者不**可能**引起全球灾难破坏的情况下起作用。对人工智能而言，这种可能性或不可能性是设计空间的自然特性 － **可能性**并不比光速或万有引力常数更多受到人们主观抉择的影响。

11：人工智能对比人类智能增强

我不认为**人类**会继续这样发展直到无穷的未来，几千年或几百万年或几十亿年以后，而期间**从来**不出现**任何**打破当代智力上限的大脑。如果是那样，总有一天人类会**首次**面对比人类更聪明的智慧的挑战。如果我们在第一轮就胜出，那么我们可以在以后的各局挑战中号召比人类更聪明的智慧来迎战。

或许我们情愿采取不同于人工智能的途径来达到比人类聪明的智能 － 例如，增加人类本身的智力？举一个极端的例子，假设有人说：人工智能的前景使我感到紧张。我宁愿，在任何人工智能开发出来之前，把单个人脑扫描入计算机，一个一个神经元地扫描，然后缓慢而确定地升级，直到它们变得超级聪明；**那**就是人类将要面对超级智慧挑战的基础。

于是我们面临两个问题：这种情况可能吗？如果可能，这种情况是人们期望的吗？（按这种顺序问这两个问题更为明智，这是出于理性的原因：我们应该避免将感情因素附加于某项实际上并不是选项的选择上。）

让我们假设一个人被逐个神经元地扫描入计算机，如莫拉维克（1998）所提议的。这就必然要求计算机的容量远远**超过**人脑的计算能力。假设，计算机可以模拟人脑的生物细节，精确避免所有可检测到的由系统性的底层错误引起的高层效应。**无论何种**生物上的意外事件以何种方式影响到信息处理过程，我们必须忠实地模拟直至精确性足以维持总体信息处理流的同构。要**模拟**像人脑这样混乱的生物计算机，我们需要比包含在人脑本身以内的更**强大**许多的计算能力。

最可能使我们有能力逐个神经元扫描大脑 – 以足够捕获**每个**认知相关的神经结构的精细度 – 的就是发明先进的分子纳米技术。分子纳米技术或许可以生产出一个总处理能力超过当前所有人类大脑思维能力总和的台式计算机。（博斯特罗姆 1998；莫拉维克 1999； 默克与卓克斯勒 1996； 沙德伯格 1999）

此外，如果技术允许我们以高保真的方式扫描大脑并把扫描作为**代码执行**，那么这就意味着在此之前的几年里，技术已足够先进到可以获得大脑神经回路信息处理的**极其精确的**图像，也许研究者早已在尽力去理解这种技术。

并且，要对上传的信息进行**升级** – 转变大脑扫描数据以增加它的智能 – 我们必须对大脑的**高级功能**有所理解，以及它们是如何促成智慧，而且达到出色的精细程度。

还有，人类不是天生就能通过外界的神经科学家或内部递归自我完善来改进的。自然选择并没有把人脑塑造成可以人为介入的。大脑中所有复杂的机械装置都只适应于较窄的参数范围。假设你能让人变得更聪明，别说是超级聪明；那么他还能保持**心智健全**吗？人脑非常容易受到干扰；仅仅改变神经递质的平衡就可以引发精神分裂，或其它紊乱。迪肯 （1997）有一篇关于人脑进化的精彩讨论文章，大脑的原件是如何精细地被平衡着，而这又如何反应在当代脑功能失常的案例中。人类的大脑不是终端用户可修改的。

所有这些都使在**任何一个人工智能被制造出来**之前首先实现**人脑扫描与升级**显得非常不可行。当能够上传大脑详细信息的技术最初实现的时候，那也意味着**前所未有的超大计算能力的出现**，以及或许是**大幅领先的认知科学的进步**，这些已远远超过创造一个人工智能所需。

从零开始造一架 747 不容易。但是相对容易的是：

- 从**生物学鸟类的现有设计**着手，
- **逐步递增式地通过一系列相继的阶段修改设计**，
- 每一个阶段都是**独立可行**的，
- 如此继续直到一只**鸟被放大成 747 的尺寸**，
- 它**确实能飞**，
- **飞得像 747 一样快**，

• 接着**把这一系列转变在一只真实的鸟身上实行，**
• **并且不把鸟杀死也没有让它觉得非常不舒服**？

我不是说这个方案永远不可能实现。我是说如果直接设计一个 747，并实现它，把它比作是鸟的升级，这样会**更容易**一些。“让我们把一个现存的小鸟放大到 747 的尺寸”并**不是**回避令人生畏的空气动力学理论的聪明手段。也许，最初你所知道有关飞行的所有知识就是鸟类拥有飞行的神秘本质，而你用来制造 747 的材料就躺在那里。但是除非飞行的秘密对你来说已完全解开，否则你无法塑造飞行的神秘本质，即使它早已存在于鸟中。

以上的论点是故意指向一个极端的例子。所要阐明的一般观点是我们没有**绝对**的自由去拾起一条听上去美好又保险的途径，或是实现那种作为科幻小说而言是不错的好故事。我们受限于可以领先他者的技术。

我不反对将人脑扫描入计算机并使它们更聪明，但这看上去非常不像是引起人类**首次**面对超出人类智慧的挑战的基础。基于能**上传和升级**人脑所需技术的各种**严格子集**，人们可以:

• 原位升级生物脑 （例如，加入有用的新神经元）；
• **或**建立人脑与计算机的实用人机接口；
• **或**建立人脑之间的实用互连接口；
• **或**构建人工智能。

此外，合理地把一个普通人的智商提高到 140 是一回事，而把一个诺贝尔奖金获得者的智力提高到超出人类又是另一回事。（暂且不提 IQ 或赢得诺奖作为动态智力衡量的可疑性；请原谅我的隐喻。）服用脑康复（或喝咖啡）可能，或不能，使某些人变得聪明；但不会使你**实质上比爱因斯坦聪明**。在那种情况下我们没有赢得有重大意义的能力；我们没有让更深一步的问题变得容易；我们没有突破已有的处理生存危机的智力上限。从对付生存危机的角度看，任何提高智力的技术如果不能产生出**比人类更聪明**的（温和，理智的）大脑，那么人们就会质疑是否将同样的金钱和时间投入到寻找一个当今世界上绝顶聪明的人，说服他就同样的问题发挥才能，会更有效呢？

而且，人脑的“自然”设计界限是大脑本身可表征的远古环境，每个大脑构件都适应了这种环境，你要是离该“自然”界限越远 ， 那么造成个体精神失常的危险就越大。如果智力扩增产生实质上超过人类的智慧，这也是一个全球灾难性风险。一个邪恶的智力扩增后的人能造成多大的破坏？哦。。。他们有多大的创造力？我头脑中闪现的第一个问题就是，“他们的创造力是否足以构建他们自己的递归自我改进人工智能？”

激进的人类智力增强技术本身就带有安全隐患。需要再次重申的是，我并非宣称智力增强工程因为这些问题的存在而不可能实现，我只是指出这些问题的存在。

人工智能有安全问题；而人类智力增强也同样有安全问题。并非每个严词厉色的都是你的敌人，也不是每个轻声细语的都是你的朋友。一方面，一个善良的人从宽广的道德，伦理，和由复杂的体系所描述的我们称为“友好的”抉择**出发**。另一方面，一个人工智能可以被**设计成**稳定的递归自我改进系统，而且符合安全性要求：自然选择在设计人脑时没有采取多重预防措施和保守的抉择过程，以及多个数量级的安全界限。

人类智力增强有权作为一个独立的问题，而不是人工智能的副主题；本章没有空间来详细讨论它。值得一提的是我在开始自己事业的早期同时考虑了人类智力增强和人工智能，并决定付诸努力在人工智能上。主要是因为我不认为**有用的，超越人类的**智力增强技术，会在递归自我该进的人工智能之前，实现并对后者产生**实质性的影响**。如果我被证明是错的，我将感到意外的惊喜。

当别人忙于实现人类智慧增强时，指望他们带来的增强人能更好解决问题，从而故意选择**搁置**友好人工智能的研究，我不认为这是一种可行的策略。我不愿接受一个如果人类智力增强比建造人工智能耗时更久就会**彻底**失败的策略。（反之亦然。）我担心从生物研究入手会花费太多的时间 - 太容易引起惰性，太多与自然选择已有的差劲设计方案相抗争的地方。我担心管理机构不会批准用人做实验。而且即使是人类的天才也要学习多年才能达到炉火纯青；如果要使智力增强后的人学得更快，那么要把人增强到那一水平也就更难。

如果增强智力的人出现，并在任何其它人有机会之前造出友好人工智能，我将会感到意外的惊喜。但那些期望看到这种结果的人可能需要付出艰苦努力以加速发展智力增强技术；减慢步伐将很难说服我。如果人工智能**天生**就比智力增强困难许多，那么没关系；如果造一架 747 **天生**就比放大一只鸟来得容易，那么问题就很严重了。刻意停止人工智能的研究在一个很小的**可能**范围内会有帮助，而在很大的可能范围内是无关或有害的。即使人类智力增强是可能的，它确实存在着困难的安全问题；我不得不严肃地考虑我们是希望友好人工智能先于智力增强实现呢，还是反过来。

我对友好人工智能比人类智力增强更容易实现或更为安全的断言，并不赋予十足的信心。有许多可能的途径增强人类智慧。也许它们中有一个技术会比人工智能更容易也更安全，同时也足够强大到可以对生存危机带来改观。如果是那样，我会换工作。但我的确希望指出一些注意事项以反对那种**毫不怀疑地假定**人类智力增强更简单，更安全，且足够强大能带来实质性变化的观点。

12：人工智能与其它技术的交互

加速想要的技术是一个局部策略，而**减缓**危险的技术却是一个困难的众数策路。**终止或放弃**不想要的技术则需要不太可能达到的全体策略。我建议，我们不是去考虑是否要开发一项技术，而是从我们**实际拥有的加速或减缓**一项技术的能力范围出发，看看在**该范围以内**，我们期望哪种技术**先于**其它被开发。

在纳米技术领域，所提出的目标常常是在进攻性的技术之前优先开发防御性的技术。我对此深表担忧，因为一个**给定级别**的攻击性技术倾向于比防御它的技术要容易许多。在人类文明史上，进攻多半是胜过防守。手枪比防弹背心的发明要早几个世纪。在天花疫苗被开发出来前天花被当成是一种战争工具。至今还没有能够抵挡核爆炸的防御系统；国家之间并非是依靠防御力实现自我保护，而是通过核威慑达到平衡。纳米技术同样具有与生俱来的难于防御的问题。

那么我们是应该期望纳米技术先于人工智能还是后于其被开发出来？如前所述，这是一个棘手的问题。答案与纳米技术或人工智能作为一种生存危机的内在困难性无关。至于它们的先后**次序**问题，我们应该问的问题是：“人工智能可以帮我们解决纳米技术问题吗？反之，纳米技术能帮我们对付人工智能吗？”

在我看来人工智能的成功解决能在很大程度上有益于我们发展纳米技术。我倒是看不出纳米技术怎样能使**友好**人工智能问题变得更简单。如果巨型纳米计算机使得开发人工智能变得更容易却**未能**使友好性这一特定问题的解决变得更明朗，那么它们之间是一种**不利的**交互。因此，在其它条件均相等的情况下，我极其主张友好人工智能**先于**纳米技术被开发。如果我们接受人工智能的挑战并取得胜利，我们可以号召一个友好人工智能来帮我们开发纳米技术。如果我们开发出纳米技术并且还活着，我们仍然需要面对人工智能的挑战。

一般说来，一个**成功的**友好人工智能应该可以帮助解决几乎所有的问题。因此，如果一种技术使得人工智能变得更容易或更难，但却带着灾难性风险，我们应该在其它条件均等的情况下**首先**面对人工智能的挑战。

任何可以增加现有计算能力的技术减少了开发人工智能所需的最低理论复杂度，但它如果并没有在**友好性**这方面带来什么帮助，我把它算作一个净负值。疯狂科学的摩尔定理：每过 18 个月，毁灭世界所需要的智商下限就会下降一个百分点。

成功实现人类智力增强会使人工智能更容易，也能帮到其它技术。但人类增强本身并**不**比人工智能更容易，更安全；而且如果它们中的一种天生就比另一种更容易的话，我们现实已有的能力范围还不能反转人类智力增强和人工智能之间的自然顺序。

13：促进友好人工智能的发展

“我们提议，1956 年夏在新罕布什尔州汉诺威市的达特茅斯学院，由 10 位科学家用 2 个月时间实现人工智能。这项研究是基于这样一个猜想：人类学习的每个方面或智慧的其它特性原则上都可以被精确描述，以至于可以由一个机器来模拟它。我们将尝试解决如何使机器使用语言，形成抽象思维和概念，解决目前仅限于人类才能解决的问题，并且它们可以自我进化。我们认为如果仔细挑选一组精英科学家来一起努力一个夏天，我们将会在这些问题中的一个或几个上实现重大进

展。”

-- 麦卡锡，明斯基，罗切斯特，和香农 （1955）。

关于人工智能的**达特茅斯暑期研究项目提议**是有史以来首次使用“人工智能”一词。他们没有先验知识警告他们这个问题是非常难的。对于他们说的，仅用一个夏天的时间，“一个重大的进展**能**被实现”而不是“**可能**被实现”，我仍然称之为一个真正的错误，那是对于问题的难度和解决时间的一个特定猜想，带有特别的不可能性。但如果他们使用“**可能**”一词，我就无法反对。他们怎么会知道呢？

达特茅斯提议包括以下和其它一些主题：语言交流，语义理解，神经网络，抽象，随机和创造性，与环境的交互，大脑模型，原创性，预见，发明，发现和自我改进。

现在从我看来一个具有语言，抽象思维，创造性，环境交互，原创性，预见性，发明，发现，最重要的是，自我改进能力的人工智能，早已**远远超出**同时具有友好性的阶段。

达特茅斯提议对制造一个善意的/美好的/仁慈的人工智能只字未提。安全问题也未提及，即使是仅仅为了提供一些可被删除的部分。这就是那时的情形，即使在人们认为达到人类水平的人工智能已触手可及的那个光明的夏天，人们都忽视了人工智能的友好性与安全性。**达特茅斯提议**是 1955 年起草的，那是在针对生物技术的阿西洛玛会议，反应停畸形儿事件，切尔诺贝利核泄漏事故以及 911 发生之前。如果今天人工智能的想法**刚刚被提出**，那么一定**有人**会要求知道相应的风险管理的特别措施。我没有说这种改变对人类的文明来说是好还是坏。我没说这会导致好还是坏的科学研究。关键是如果**达特茅斯提议**是晚了 50 年之后提出的，安全性一定是主题之一。

当我 2006 年写下此文时，人工智能研究社群仍没有看出友好人工智能是问题的一部分。我希望能就此引一些文献，但它们却并不存在。友好人工智能没有进入人们的**概念**空间，它不仅仅是不流行或是缺乏资助。你甚至不能称友好人工智能为地图上的空白一点，因为人们甚至没有意识到它的缺失。如果你读过一些流行的/半技术化的提议如何造人工智能的书籍，譬如《哥德尔，埃舍尔，巴赫》（霍夫施塔特 1979）或是《智力社会》（明斯基 1986），回想一下，你可能没有印象友好人工智能出现在关于人工智能挑战性的讨论中。我也没看见过友好人工智能作为一个技术问题在任何技术文献中被讨论过。我在文献搜索上的尝试只给出一些简短的非技术文章，彼此并无连贯，除了阿莎科·阿西莫夫的“机器人三定理”（阿西莫夫 1942）之外，没有共同引用的重要文献。考虑到现在是 2006 年，为什么没有更多的人工智能学者谈论安全性问题？我无权探知他人的心理，但我会就一些个人讨论的结果作个简要的推断。

人工智能领域已经适应了它在过去 50 年的经历，尤其是这种模式：即一开始许诺

达到人类能力的雄心壮志，随之而来的便是令人尴尬的公开失败。把这种尴尬归咎于“人工智能”也许是不公平的；新闻是不会宣传明智而没有夸下海口的研究者的保守言论的。当先进的人工智能被提及时，即刻呈现在人们脑海中仍然是那些失败了的高谈阔论，这在人工智能研究领域的内外都一样。人工智能的研究文化已经适应了这一情形：谈论类人的能力是禁忌的。还有一个更强的禁忌针对任何宣称或预测某种能力而却没有相应的可运行代码加以演示的行为。我遇到一种观点认为**任何号称在从事友好人工智能研究的人，都暗指他们的人工智能设计足够强大到可以满足友好性的需要**。

显然这个观点既不是逻辑上的真言，也非现实中的明理。如果我们想象有人创造了一个真实的，成熟的人工智能，它足够强大到**必须是友好的**，而且，正如我们期望的，它也**确实是友好的**，那么，创造它的人一定是在友好人工智能研究上花费了多年心血。友好人工智能不是一个你可以在当它首次被需要的时候实时发明出来的模块，可以拴在已有的，优美的设计之上而不对后者产生任何改变。

人工智能领域有一些技术，例如神经网络和进化程序设计，它们随着几十年的慢慢调整已日渐强壮。但神经网络是不透明的 – 用户不知道网络是如何做出抉择的 – 它也不容易提供透明度；发明和精化神经网络的人们没有考虑友好人工智能的长期问题。进化编程（EP）是随机的，它不精确保存所产生代码中的优化目标；EP 多半能在测试过的环境中给出满足你所需功能的代码，但那些代码可能还会做一些其它事情。EP 是一个强大的，正在慢慢成熟的，**本质上**不适合友好人工智能需求的技术。友好人工智能，正如我所提议的，需要递归自我改进的重复循环从而能够保持一个稳定的优化目标。

目前最强大的人工智能技术，虽然它们随着时间经过了发展，精化和改进，就我看来它们与友好人工智能的需要基本不兼容。千年虫问题 – 尽管没有造成全球灾难却耗费了昂贵的修复代价 – 是没能预见到未来设计需要的典型案例。最可怕的情形是我们发现自己陷在成熟的，强大的，公开可用的人工智能技术里，而它们结合起来却产生出**非友好的**人工智能，并**不能**用来制造友好人工智能，除非之前 30 年人工智能的研究全部推翻重来。

在人工智能领域里，公开讨论**达到人类能力的**人工智能是需要勇气的，鉴于这个领域里此种讨论的过往经历。富于诱惑的行为是祝贺自己如此勇敢竟然仍能对其加以讨论，并且知道适时而止。在已经如此勇敢之后，还要讨论**超人类**人工智能看上去似乎是荒谬和完全没有必要的。（但是没有特别的原因说明为什么人工智能会历经艰辛在智力的标尺上蹒跚而上，而却永远停在人类这一点上。）敢于谈论**友好人工智能**，作为对**超人类**人工智能的全球灾难性风险的预警，比勇于显示自己的超凡见识和过人胆量需要的勇气级别还要高出两级。

有一种实用主义的反对观点承认友好人工智能是一个重要的问题，但是担心，鉴于我们目前的理解力，我们还不处于解决友好人工智能的阶段：如果我们**现在**就试着去解决这个问题，我们只会失败，或是从事反科学而不是科学。

这个反对观点是值得担忧的。在我看来知识已经在那儿了 – 有可能通过学习现有知识中充分大的部分，进而着手解决友好人工智能问题而不至于开始就一头撞在墙上 – 但是相关的知识散落在**多个学科领域**：抉择理论**和**进化心理学**和**概率论**和**进化生物学**和**认知心理学**和**信息论**以及**传统的“人工智能”。。。没有现成的课程为推动友好人工智能的研究进展准备好一大批现成的研究者。

天才的“十年法则”，说的是没有人能在哪个领域里获得突出的成就而不付出起码十年的努力，已被从数学到音乐到竞争性网球等的各个领域证实。（海耶斯 1981）莫扎特 4 岁开始作交响曲，但它们还不是**莫扎特**交响曲 – 莫扎特又花了 13 年时间才开始编写**出色的**交响曲。（韦斯伯格 1986）我自己对于学习曲线的亲身经历强化了这一担忧。如果我们希望人们能够在友好人工智能上取得进展，那么他们不得不开始培训自己，全时制的，在他们**急切被需要**的多年之前。

如果明天比尔和美琳达•盖茨基金分配一亿资金用于友好人工智能的研究，那么数以千计的科学家立刻就会开始写项目申请书并使之与友好人工智能显得密切相关。但他们不是真正地对这个问题**感兴趣** – 证据就是他们在有人提供资金之前没有表现出好奇心。当通用人工智能不合时尚而友好人工智能完全不在大众视线扫描范围内时，我们至少可以假设任何谈及这个问题的人是对它真正有兴趣的。如果你砸太多钱在一个某领域尚未准备好解决的问题上时，过量的资金更可能产生的是反科学 – 也即一大团错误的解决方案 - 而不是科学。

我无法把这个结论当作是好消息。如果友好人工智能可以通过投入大量人力和财力得到解决，我们全都会更加安全。但就目前也即 2006 年看来，我非常怀疑这种可能性 – 友好人工智能领域，包括人工智能领域本身，正处于极度混乱状态中。然而要是有人断言我们现在还知之甚少，我们**没有能力**在友好人工智能方面取得进展，那么我们应该反问他在下此结论之前花了多少时间进行学习。谁能说什么是科学所**不**知道的？对任何一个人而言都有太多知识要学。在没有学遍所有新奇的领域之前，谁能说我们还**没有**准备好解决一个科学问题？并且如果我们因为尚未准备充分而**没有能**在友好人工智能上取得进展，那么这并不意味着我们不**需要**友好人工智能。这两句话**完全是**不同的！

因此如果我们发现我们**没能**在友好人工智能上有所进展，那么我们就要尽快找出原因，走出困境，越快越好！无论如何都没有哪种保证说，仅仅因为我们无法控制一个危机，它就会离我们而去了。

假如有尚未建树的优秀青年科学家们自发地对友好人工智能感兴趣，那么我认为如果他们能申请到一个多年期的项目经费全职从事该问题研究的话，将会对人类带来很大的益处。为此友好人工智能需要一些经费资助 – 远比现有的要多。不过我担心在此开始阶段，一个曼哈顿计划规模的项目，比起信号只会增加更多的噪音。

结束语

我一度认为当代文明处于一种非稳定状态。I.J.古德的智力爆炸假说描述了一个动态不稳定系统，就像是一只钢笔在它的尖端上竖起。如果那只钢笔是**恰好**垂直于桌面的，它可以保持直立；但只要钢笔稍微倾斜一丁点，重力会往其倾斜方向进一步拉倒它，而且这个过程会以一定加速度发生。类似的，一个有智能的系统会在某一时刻加速自身的智力。

一颗无生命的行星，绕着它的恒星运行，也是稳定的。智力爆炸的情况则不同，灭亡不是一个**动态**吸引子 – 在**几乎**灭绝和彻底灭绝之间有巨大的差距。尽管如此，**完全的**灭绝是稳态的。

我们的文明是否终将步入这两种稳态中的一种？

从逻辑角度，上述论点是有漏洞的。想想大蛋糕谬误，例如：智力不是盲目地移进吸引子的，它们要有动机。即便如此，我怀疑，说句**实**话，我们的出路将归结为变得更加聪明或是灭绝。

自然不是冷酷无情，而是平衡中立的；中立得常常使人们觉得它与完全的敌意无甚差别。现实将一个又一个挑战抛向你，而当你遇到一个解决不了的难题时，你却要自己承受后果。自然常常提出非常不公平的要求，即使在那些失败的惩罚是死亡的测试中也如此。一个中世纪的农民该如何发明治疗肺结核的方法？自然并不会使她的挑战符合你的技能，或你的资源，或是你有多少空余时间来思考她的问题。于是当你撞上一个对你而言太难无法解决的致死的难题时，你死路一条。也许那样想会使人不愉快，但它却是人类成千上万年来的真实经历。同样的事情也会很容易发生在整个人类身上，如果人类撞上了一个不公平的挑战。

如果人类不会衰老，从而百岁老人和 15 岁少年有同样的死亡率，我们不会永生。我们只会存活直到死亡的概率赶上我们。即使只要活上一百万年，作为一个不会衰老的人活在像我们的世界一样高风险的环境中，你必须以某种方式把你的年度事故发生率降到接近于**零**。你或许不能开车；不能乘飞机；你或许不能横穿马路即使你已环顾左右，因为它还是太危险了。即使你丢舍所有趣事，放弃所有生活内容，你也不能跨越障碍活到一百万年。那并非物理上不可能，而是**认知上**不可能。

人类，即使不会衰老但也不会永生。人类能存活至今仅仅是因为，在过去的几百万年间，没有装载氢弹的军械库，没有控制小行星飞向地球的宇宙飞船，没有生物武器实验室生产超病毒，没有重复出现在年度展望中的核战争或纳米技术战争或凶猛的人工智能。要想继续存活相当长的时间，我们需要把**每个**风险都降至接近**零点**。“相当好”不足以好到持续未来的一百万年。

看上去似乎有些不公平。控制风险的能力有史以来不是人类体系的强项，不管人

们多么努力。几十年来美国和俄罗斯尽量避免核战，但不**完美**；有时核战一触即发，例如在 1962 年古巴导弹危机中。如果我们假定未来的人类智力同样兼有愚蠢和智慧，同样英雄主义与自我中心并存，如同我们在历史书中读到的一样 – 那么生存危机的游戏早就结束了；我们一开始就输了。我们可能再活十年，甚至是一个世纪，但不会是一百万年。

但是人类的大脑不是可能性的极限。**人类**代表**第一个**通用智慧。我们生而处于一切的开端，智慧的伊始。幸运的话，未来的历史学家回过头来看时会将我们当前的世界描述成为一个处于中间阶段的笨拙的青少年时期，一个人类聪明到足以给自己创造出惊人难题，却还没有聪明到足以解决它的时期。

然而在我们还没有走出青春期的时候，我们必须，作为青少年，面对一个成年人的问题：超越人类智慧的难题。这是我们走出生命周期高死亡率阶段的途径；是关闭我们脆弱之窗的方法；它或许也是我们面临的最最危险的那一个危机。人工智能是迎接挑战的一条出路；而且我认为最终我们会选择这条路。我认为，最后我们将得到证明，从头开始造一架 747 要比把现有的小鸟扩大或是给它装上喷气式引擎要容易得多。

我不想贬低那些尝试按照精确设计和目标制造比我们聪明的智能体的大胆设想。只是让我们停下来回想一下，**智慧**并非是第一个人类科学曾经遇到过的极难理解的问题。星系曾经也是神秘的迷团，还有化学，生物。一代又一代的研究者尝试理解这些谜题却以失败告终，于是它们便有了不可能被纯科学方法解决的名声。很久以前，没有人明白为什么一些物质是惰性和没有生命的，而其它物质却充满勃勃生机。没有人知道生命是如何实现自我复制，以及我们的双手为何服从头脑的指挥。开尔文勋爵写道：

> “动物或植物体对物质世界的影响是完全超乎任何科学所能开始探究的范围之外的。生命体控制移动粒子的能力，展现于我们人类自由意志每天所创造的奇迹中，成长在植物从一个种子开始世代相传的生长历程里，它与任何可能的原子的偶然并行作用的结果是截然不同的” （麦克菲 1912 引文）

所有的科学上的无知都被古人神话了。每一项知识的空缺都可以追溯到人类好奇心的最初；那些知识漏洞历经岁月洗礼，看上去是永恒存在的，直到有人把它们补上。我认为仅仅凭易犯错的人类就能成功战胜友好人工智能挑战的可能性是存在的。但我们首先必须不再把智慧当成神圣的秘密，如同开尔文对生命的看法。智慧必须不再是任何种类的秘密，无论神圣与否。我们必须像是开始一门真正的艺术一样着手创造人工智能。那样也许我们会成功。